Complementarity and Aggregate Implications of Assortative Matching: A Nonparametric Analysis*

Bryan S. Graham† Guido W. Imbens‡ Geert Ridder§

First Draft: July 2004
This version: March 2006

Abstract

This paper presents methods for evaluating the effects of reallocating an indivisible input across production units. When production technology is nonseparable such reallocations, although leaving the marginal distribution of the reallocated input unchanged by construction, may nonetheless alter average output. Examples include reallocations of teachers across classrooms composed of students of varying mean ability and altering assignment mechanisms for college roommates in the presence of social interactions. We focus on the effects of reallocating one input while holding the assignment of another, potentially complementary input, fixed. We present a class of such reallocations – correlated matching rules – that includes the status quo allocation, a random allocation, and both the perfect positive and negative assortative matching allocations as special cases. Our econometric approach involves first nonparametrically estimating the production function and then averaging this function over the distribution of inputs induced by the new assignment rule. Formally our methods build upon the partial mean literature (e.g., Newey 1994, Linton and Nielsen 1995). We derive the large sample properties of our proposed estimators and assess their small sample properties via a limited set of Monte Carlo experiments. An application, assessing the effects of spousal sorting on child education (e.g., Kremer 1996), concretely illustrates our methods.

JEL Classification: C14, C21, C52

Keywords: *Average Treatment Effects, Complementarity, Aggregate Redistributional Effects*

*Financial support for this research was generously provided through NSF grant SES 0136789 and SES 0452590. We thank participants in the Harvard-MIT and Brown Econometrics Seminars for comments. We thank Cristine Pinto for excellent research assistance.

†Department of Economics, University of California at Berkeley, 665 Evans Hall, Berkeley, CA 94720-3880. E-Mail: bgraham@econ.berkeley.edu, Web: http://www.econ.berkeley.edu/~bgraham/

‡Department of Agricultural and Resource Economics and Department of Economics, University of California at Berkeley, 661 Evans Hall, Berkeley, CA 94720-3880, and NBER. E-Mail: imbens@econ.berkeley.edu, Web:http://elsa.berkeley.edu/users/imbens/.

§Department of Economics, University of Southern California, 310A Kaprielian Hall, Los Angeles, CA 90089. E-Mail: ridder@usc.edu, Web: http://www-rcf.usc.edu/~ridder/

1 Introduction

Consider an input into a production process. For each firm output may be monotone in this input, but at different rates. If the input is indivisible and its aggregate stock fixed, it will be impossible to simultaneously raise the input level for all firms. In such cases it may be of interest to consider the output effects of *reallocations* of the input across firms. Here we investigate econometric methods for assessing the effect on average output of such reallocations. A key feature of reallocations is that while potentially altering input levels for each firm, they keep the marginal distribution of the input across the population of firms fixed.

We consider a two parameter family of feasible reallocations that include several focal allocations as special cases. Reallocations in this family may depend on the distribution of a second input or firm characteristic. This characteristic may be correlated with the firm-specific return to the input to be reallocated.

One reallocation redistributes the input across firms such that it has perfect rank correlation with the second input. We call this allocation the positive assortative matching allocation. We also consider a negative assortative matching allocation where the input is redistributed to have perfect negative rank correlation with the second input. A third allocation involves randomly assigning the input across firms. This allocation, by construction, ensures independence of the two inputs. A fourth allocation simply maintains the status quo assignment of the input.

Our family of reallocations, which we call correlated matching rules, includes each of the above allocations as special cases. In particular the family traces a path from the positive to negative assortative matching allocations. Each reallocation along this path keeps the marginal distribution of the two inputs fixed, but is associated with a different level of correlation between the two inputs. Each of the reallocations we consider are members of a general class of reallocation rules that keep the marginal distributions of the two inputs fixed.

We derive an estimator for average output under correlated matching. Our estimator requires that the first input is exogenous conditional on the second input and additional firm characteristics. Except for the case of perfect negative and positive rank correlation the estimator has the usual parametric convergence rate. For the two extremes the rate of convergence is slower. In all cases we drive the asymptotic distribution of the estimator.

Our focus on reallocation rules that keep the marginal distribution of the inputs fixed is appropriate in applications where the input is indivisible, such as in the allocation of teachers to classes or managers to production units. In other settings it may be more appropriate to consider allocation rules that leave the total amount of the input constant by fixing its average level. Such rules would require some modification of the methods considered in this paper.

Our methods may be useful in a variety of settings. One class of examples concerns complementarity of inputs in production functions (e.g. Athey and Stern, 1998). If the first and second inputs are everywhere complements, then the difference in average output between the positive and negative assortative matching allocations provides a nonparametric measure of the degree of complementarity. This measure is invariant to monotone transformations of the inputs. If the production function is not supermodular interpretation of this difference is not straightforward, although it still might be viewed as some sort of 'global' measure of input complementarity. With this concern in mind we also provide a local measure of complementarity. In particular we consider whether small steps away from the status quo and toward the perfect assortative matching allocation raise average output.

A second example concerns educational production functions. Card and Krueger (1992) study the relation between educational output as measured by test scores and teacher quality. Teacher quality may improve test scores for all students, but average test scores may be higher or lower depending on whether, given a fixed supply of teachers, the best teachers are assigned to the least prepared students or vice versa. Parents concerned solely with outcomes for their own children may be most interested in the effect of raising teacher quality on expected scores. A school board, however, may be more interested in maximizing expected test scores given a fixed set of classes and teachers by optimally matching teachers to classes.

A third class of examples arises in settings with social interaction (c.f., Manski 1993; Brock and Durlauf 2001). Sacerdote (2001) studies the peer effects in college by looking at the relation between outcomes and roommate characteristics. From the perspective of the individual student or her parents it may again be of interest whether a roommate with different characteristics would, in expectation, lead to a different outcome. This is what Manski (1993) calls an exogenous or contextual effect. The college, however, may be interested in a different effect, namely the effect on average outcomes of changing the procedures for assigning roommates. While it may be very difficult for a college to change the distribution of characteristics in the incoming classes, it may be possible to change the way roommates are assigned. In Graham, Imbens and Ridder (2006b) we consider average effect of segregation policies.

In all these cases we focus on policies that change the way a fixed distribution of inputs is allocated to a population of units with a fixed distribution of characteristics. We are interested in the effect such policies have on the distribution of outcomes. Typically the most interesting measure will be the average level of the outcome. We will call the causal effects of such policies Aggregate Redistributional Effects (AREs).

If production functions are additive in inputs the questions posed above have simple answers: average outcomes are invariant to input reallocations. While reallocations may raise individual outcomes for some units, they will necessarily lower them by an offsetting amount for others. Reallocations are zero-sum games. With additive and linear functions even more general assignment rules that allow the marginal input distribution to change while keeping its average level unchanged do not affect average outcomes. In order for these questions to have interesting answers, one therefore needs to explicitly recognize and allow for non-additivity and non-linearity of a production function in its inputs. For this reason our approach is fully nonparametric.

The current paper builds on the larger treatment effect and program evaluation literature.[1] More directly, it is complementary to the small literature on the effect of treatment assignment rules (Manski, 2004; Dehejia, 2004; Hirano and Porter, 2005). Our focus is different from that in the Manski, Dehejia, and Hirano-Porter papers. First, we allow for continuous rather than discrete or binary treatments. Second, our assignment policies to do not change the marginal distribution of the treatment, whereas in the previous papers treatment assignment for one unit is not restricted by assignment for other units. Our policies are fundamentally redistributions. In the current paper we focus on estimation and inference for specific assignment rules. It is also interesting to consider optimal rules as in Manski, Dehejia and Hirano-Porter. The class of feasible reallocations/redistributions includes all joint distributions of the two inputs with fixed marginal distributions. When the inputs are continuously-valued, as we assume in the current paper, the class potential rules is very large. Characterizing the optimal allocation within this class is therefore a non-trivial problem. When both inputs are discrete-valued the problem of finding the optimal allocation is tractable as the joint distribution of

[1] For recent surveys see Angrist and Krueger (2001), Heckman, Lalonde and Smith (2001), and Imbens (2004).

the inputs is characterized by a finite number of parameters. In Graham, Imbens and Ridder (2006a) we consider optimal allocation rules when both inputs are binary, allowing for general complementarity or substitutability between the two inputs.

Our paper is also related to recent work on identification and estimation of models of social interactions (e.g., Manski 1993, Brock and Durlauf 2001). We do not focus on directly characterizing the within-group structure of social interactions, an important theme of this literature. Rather our goal is simply to estimate the average relationship between group composition and outcomes. The average we estimate may reflect endogenous behavioral responses by agents to changes in group composition, or even equal an average over multiple equilibria. Viewed in this light our approach is reduced form in nature. However it is sufficient for, say, an university administrator to characterize the outcome effects of alternative roommate assignment procedures.

The econometric approach taken here builds on the partial mean literature (e.g., Newey, 1994; Linton and Nielsen, 1995). In this literature one first estimates a regression function nonparametrically. In the second stage the regression function is averaged, possibly after some weighting with a known or estimable weight function, over some of the regressors. Similarly here we first estimate a nonparametric regression function of the outcome on the input and other characteristics. In the second stage the averaging is over the distribution of the regressors induced by the new assignment rule. This typically involves the original marginal distribution of some of the regressors, but a different conditional distribution for others. Complications arise because this conditional covariate distribution may be degenerate, which will affect the rate of convergence for the estimator. In addition the conditional covariate distribution itself may require nonparametric estimation through its dependence on the assignment rule. For the policies we consider the assignment rule will involve distribution functions and their inverses similar to the way these enter in the changes-in-changes model of Athey and Imbens (2005).

The next section lays out our basic model and approach to identification. Section 3 then defines and motivates the estimands we seek to estimate. Section 4 presents of our estimators, and derives their large-sample properties, for the case where inputs are continuously-valued. Section 5 presents a simple test for the efficiency of the status quo allocation of inputs. Section 6 deals with estimation and inference in the case where inputs take on discrete values. In this case the problem is fully parametric and large sample standard errors can be computed using the delta method Section 7 presents an application and the results of a small Monte Carlo exercise.

2 Model

In this section we present the basic model and identifying assumptions. For clarity of exposition we use the production function terminology; although our methods are appropriate for a wide range of applications as emphasized in the introduction. Let $Y_i(w)$ be the output associated with input level w for firm $i = 1, \ldots, N$. We are interested in reallocating the input W across firms. We focus upon reallocations which hold the marginal distribution of W fixed. As such they are appropriate for settings where W is a plausibly indivisible input, such as a manager or teacher with a certain level of experience and expertise. The presumption is also that the aggregate stock of W is difficult to augment.

In addition to W there are two other (observed) firm characteristics that may affect output: X and Z, where X is a scalar and Z is a vector of dimension K. The first characteristic

X could be a measure of, say, the quality of the long-run capital stock, with Z being other characteristics of the firm such as location and age. These characteristics may themselves be inputs that can be varied, but this is not necessary for the arguments that follow. In particular the unconfoundedness or exogeneity assumption that we make for the first input need not hold for these characteristics.

We observe for each firm $i = 1, \ldots, N$ the level of the input, W_i, the characteristics X_i and Z_i, and the realized output level, $Y_i = Y_i(W_i)$. In the educational example the unit of observation would be a classroom. The variable input W would be teacher quality, and X would be a measure of quality of the class, e.g., average test scores in prior years. The second characteristic Z could include other measures of the class, e.g., its age or gender composition, as elements. In the roommate example the unit would be the individual, with W the quality of the roommate (measured by, for example, a high school test score), and the characteristic X would be own quality. The second set of characteristics Z could be other characteristics of the dorm or of either of the two roommates such as smoking habits (which may be used by university administrators in the assignment of roommates).

Our identifying assumption is that conditional on firm characteristics $(X, Z')'$ the assignment of W, the level of the input to be reallocated, is unconfounded or exogenous.

Assumption 2.1 (Unconfoundedness/Exogeneity)

$$Y(w) \perp W \,\Big|\, X, Z, \quad \textit{for all } w \in \mathcal{W} \subset \Re^1.$$

This type of assumption is common in the (binary) treatment effect literature where its precise form is due to Rosenbaum and Rubin (1983). To interpret the assumption, consider first the case where there are no additional characteristics (i.e., no $\dim(X) = \dim(Z) = 0$). Then Assumption 2.1 requires that $Y(w) \perp W$. This implies that the average output we would observe if all firms were assigned input level $W = w$ equals the average output among firms that were in fact assigned input level $W = w$

$$\mathbb{E}[Y(w)] = \mathbb{E}[Y|W = w].$$

This requires that the distribution of unobservables, or potential outcomes, for the subpopulation of firms that were assigned $W = w$ be the same as that for the overall population of firms; a condition that holds under random assignment of W.

The full assumption requires this equality to hold only in subpopulations homogenous in X and Z. Let

$$g(w, x, z) = \mathbb{E}[Y|W = w, X = x, Z = z],$$

denote the average output associated with input level w and characteristics x and z. Under unconfoundedness we have – among firms with identical values of X and Z – an equality between the counterfactual average output that we would observe if all firms in this subpopulation were assigned $W = w$, and the average output we observe for the subset of firms within this subpopulation that are in fact assigned $W = w$. That is

$$g(w, x, z) = \mathbb{E}[Y(w)|X = x, Z = z].$$

Assumption 2.1 has proved controversial (c.f., Imbens 2004). It holds under conditional random assignment of W to units; as would occur in an explicit experiment. However randomized allocation mechanisms are also used by administrators in some institutional settings. For example some universities match freshman roommates randomly conditional on responses in a housing questionnaire (e.g., Sacerdote 2001). This assignment mechanism is consistent with Assumption 2.1. In other settings, particularly where assignment is bureaucratic, as may be true in some educational settings, a plausible set of conditioning variables may be available. In this paper we focus upon identification and estimation under Assumption 2.1. In principle, however, the methods could be extended to accommodate other approaches to identification based upon, for example, instrumental variables.

Much of the treatment effect literature (e.g., Angrist and Krueger, 2000; Heckman, Lalonde and Smith, 2000; Manski, 1990; Imbens, 2004) has focused on the average effect of an increase in the value of the treatment. In particular, in the binary treatment case ($w \in \{0, 1\}$) interest has centered on the average treatment effect

$$\mathbb{E}_{X,Z}[g(1, X, Z) - g(0, X, Z)].$$

With continuous inputs one may be interested in the full average output function $g(w, x, z)$ (Imbens, 2000; Flores, 2005) or in its derivative with respect to the input,

$$\frac{\partial g}{\partial w}(w, x, z),$$

either at a point or averaged over some distribution of inputs and characteristics (e.g., Powell, Stock and Stoker, 1989; Hardle and Stoker, 1989).

Here we are interested in a different estimand. We focus on policies that redistribute the input W according to a rule based on the X characteristic of the unit. For example upon assignment mechanisms that match teachers of varying experience to classes of students based on their mean ability. One might assign those teachers with the most experience (highest values of W) to those classrooms with the highest ability students (highest values of X) and so on. In that case average outcomes would reflect perfect rank correlation between W and X. Alternatively, we could be interested in the average outcome if we were to assign W to be negatively perfectly rank correlated with X. A third possibility is to assign W so that it is independent of X. We are interested in the effect of such policies on the average value of the output. We refer to such effects as Aggregate Redistributional Effects (AREs).

The above reallocations are a special case of a general set of reallocation rules that fix the marginal distributions of W and X, but allow for correlation in their joint distribution. For perfect assortative matching the correlation is 1, for negative perfect assortative matching -1, and for random allocation 0. By using a bivariate normal cupola we can trace out the path between these extremes.

We wish to emphasize that there are at least two limitations to our approach. First, we focus on comparing specific assignment rules, rather than searching for the optimal assignment rule within a class. The latter problem is a particularly demanding in the current setting with continuously-valued inputs as the optimal assignment for each unit depends both on the characteristics of that unit as well as on the marginal distribution of characteristics in the population. When the inputs are discrete-valued both the problems of inference for a specific rule as well as the problem of finding the optimal rule become considerably more tractable. In that case any rule, corresponding to a joint distribution of the inputs, is characterized by a finite number of parameters. Maximizing estimated average output over all rules evaluated will then generally lead to the optimal rule. Graham, Imbens and Ridder (2006a) provide a detailed discussion for the binary case.

A second limitation is that of this class of assignment rules leaves the marginal distribution of inputs unchanged. This latter restriction is perfectly appropriate in cases where the inputs are indivisible, as, for example, in the social interactions and educational examples. In other cases one need not be restricted to such assignment rules. A richer class of estimands would allow for assignment rules that maintain some aspects of the marginal distribution of inputs but not others. A particularly interesting class consists of assignment rules that maintain the average (and thus total) level of the input, but allow for its arbitrary distribution across units. This can be interpreted as assignment rules that 'balance the budget'. In such cases one might assign the maximum level of the input to some subpopulation and the minimum level of the input to the remainder of the population. Finally, one may wish to consider arbitrary decision rules where each unit can be assigned any level of the input within a set. In that case interesting questions include both the optimal assignment rule as a function of unit-level characteristics as well as average outcomes of specific assignment rules. In the binary treatment case such problems have been studied by Dehejia (2005), Manski (2004), and Hirano and Porter (2005).

We consider the following four estimands that include four benchmark assignment rules. All leave the marginal distribution of inputs unchanged. This obviously does not exhaust the possibilities within this class. Many other assignment rules are possible, with corresponding estimands. However, the estimands we consider here include focal assignments, indicate of the range of possibilities, and capture many of the methodological issues involved.

3 Target estimands

The first estimand we consider is expected average outcome given perfect assortative matching of W on X conditional on Z:

$$\beta^{\text{pam}} = \mathbb{E}[g(F_{W|Z}^{-1}(F_{X|Z}(X|Z)|Z), X, Z)], \tag{3.1}$$

where $F_{X|Z}(X|Z)$ denotes the conditional CDF of X given Z and $F_{W|Z}^{-1}(p|Z)$ is the quantile of order $p \in [0,1]$ associated with the conditional distribution of W given Z (i.e., $F_{W|Z}^{-1}(p|Z)$ is a conditional quantile function). Therefore $F_{W|Z}^{-1}(F_{X|Z}(X|Z)|Z)$ computes a unit's location on the conditional CDF of X given Z and reassigns it the corresponding quantile of the conditional distribution of W given Z. Thus among units with the same realization of Z, those with the highest value of X are reassigned the highest value of W and so on.

The focus on reallocations within subpopulations defined by Z, as opposed to population-wide reallocations, is because the average outcome effects of such reallocations solely reflect complementarity or substitutability between W and X.

To see why this is the case consider the alternative estimand

$$\beta^{\text{pam2}} = \mathbb{E}\left[g\left(F_W^{-1}(F_X(X), X, Z\right)\right]. \tag{3.2}$$

This gives average output associated with population-wide perfect assortative matching of W on X. If, for example, X and Z are correlated, then this reallocation, in addition to altering the joint distribution of W and X, will alter the joint distribution of W and Z. Say Z is also a scalar and is positively correlated with X. Population-wide positive assortative matching will induce perfect rank correlated between W and X, but it will also increase the degree of correlation between W and Z. This complicates interpretation when $g(w,x,z)$ may be non-separable in w and z as well as w and x.

An example helps to clarify the issues involved. Let W denote an observable measure of teacher quality, X mean (beginning-of-year) achievement in a classroom, and Z the fraction of the classroom that is female. If begining-of-year achievement varies with gender, then X and Z will be correlated. A reallocation that assigns high quality teachers to high achievement classrooms, will also tend to assign such teachers to classrooms will an above average fraction of females. Average achievement increases observed after implementing such a reallocation may reflect complementarity between teacher quality and begining-of-year student achievement or it may be that the effects of changes in teacher quality vary with gender and that, conditional on gender, their is no complementarity between teacher quality and achievement. By focusing on reallocations of teachers across classrooms with similar gender mixes, but varying baseline achievement, (3.1) provides a more direct avenue to learning about complementarity.[2]

Both (3.1) and (3.2) may be policy relevant, depending on the circumstances, and both are identified under Assumption 2.1. Under the additional assumption that

$$g(w,x,z) = g_1(w,x) + g_2(z),$$

the estimands, while associated with different reallocations, also have the same basic interpretation. Here we nonetheless focus upon (3.1), although all of our results extend naturally and directly to (3.2).

Our second estimand is the expected average outcome given negative assortative matching:

$$\beta^{\text{nam}} = \mathbb{E}[g(F_{W|Z}^{-1}(1 - F_{X|Z}(X|Z)|Z), X, Z)]. \tag{3.3}$$

If, within subpopulations homogenous in Z, W and X are everywhere complements, then the difference $\beta^{\text{pam}} - \beta^{\text{nam}}$ provides a measure of the strength input complementarity. When $g(\cdot)$ is not supermodular interpretation of this difference is not straightforward. In Section 5 below we present a measure of 'local' (relative to the status quo allocation) complementarity between X and W.

Average output under the *status quo* allocation is given by

$$\beta^{\text{sq}} = \mathbb{E}[Y] = \mathbb{E}[g(W,X,Z)],$$

while average output under the random matching allocation is given by

$$\beta^{\text{rm}} = \int_z \left[\int_x \int_w g(w,x,z) \mathrm{d}F_{W|Z}(w|z) \mathrm{d}F_{X|Z}(x|z) \right] \mathrm{d}F_Z(z).$$

This last estimand gives average output when W and X are independently assigned within subpopulations.

The perfect positive and negative assortative allocations are focal allocations, being emphasized in theoretical research (e.g., Becker and Murphy 2000). The status quo and random matching allocations are similarly natural benchmarks. However these allocations are just four among the class of feasible allocations. This class is comprised of all joint distributions of inputs consistent with fixed marginal distributions (within subpopulations homogenous in Z). As noted in the introduction, if the inputs are continuously distributed this class of joint distributions is very large. For this reason we only consider a subset of these joint distributions.

[2] We make the connection to complementarity more explicit in Section 5 below.

To be specific, we concentrate on a two-parameter subset of the feasible allocations that have as special cases the negative and positive assortative matching allocations, the independent allocation, and the status quo allocation. By changing the two parameters we trace out a 'path' in two directions: further from or closer to the status quo allocation, and further from or closer to the perfect sorting allocations. Borrowing a term from the literature on cupolas, we call this class of feasible allocations comprehensive, because it contains all four focal allocations as a special case.

Average output under the correlated matching allocation is given by

$$\beta^{\mathrm{cm}}(\rho,\tau) = \tau\cdot\mathbb{E}[Y]+(1-\tau)\cdot\int g(w,x,z)\mathrm{d}\Phi\left(\Phi^{-1}(F_{W|Z}(w|z)),\Phi^{-1}(F_{X|Z}(x|z));\rho\right)F_Z(z), \quad (3.4)$$

for $\tau\in[0,1]$ and $\rho\in(-1,1)$.

The case with $\tau=1$ corresponds to the *status quo*:

$$\beta^{\mathrm{sq}} = \beta^{\mathrm{cm}}(\rho,1)$$

The case with $\tau=\rho=0$ corresponds to random allocation of inputs within sub-populations defined by Z:

$$\beta^{\mathrm{rm}} = \beta^{\mathrm{cm}}(0,0) = \int_z\left[\int_x\int_w g(w,x,z)\mathrm{d}F_{W|Z}(w|z)\mathrm{d}F_{X|Z}(x|z)\right]\mathrm{d}F_Z(z).$$

While the cases with $\tau=0$ and $\rho\to 1$ and -1 correspond respectively to the perfect positive and negative assortative matching allocations. More generally, with $\tau=0$ we allocate the inputs using a normal copula in a way that allows for arbitrary correlation between W and X indexed by the parameter ρ. In principle we could use other copulas.

4 Estimation and inference with continuously-valued inputs

In this section we present feasible estimators for the perfect positive, negative and correlated matching estimands and state their large sample properties. While the $\beta^{\mathrm{cm}}(\rho,\tau)$ can be estimated at standard parametric rates for $\tau\in[0,1]$ and $\rho\in(-1,1)$. Average output under the perfect positive and negative assortative matching allocations, β^{pam} and β^{nam}, can only be estimated an nonparametric rates. We therefore consider inference separately for the two cases.

4.1 Estimation of $\widehat{\beta}^{\mathrm{pam}}$, $\widehat{\beta}^{\mathrm{nam}}$ and $\beta^{\mathrm{cm}}(\rho,\tau)$

[NOTE: There is a disjunct between what follows and what was stated in Section 3 above. In particular we need to use estimators of the conditional CDFs $F_{X|Z}(X|Z)$ and $F_{W|Z}(W|Z)$ as well as the conditional quantile function $F_{W|Z}^{-1}(p|Z)$. There are several off the shelf possibilities here. For now I have proceeded with using the unconditional estimators as before. Thus these estimators calculate average output under "population wide" reallocations not realloactions within subpopulations homogenous in Z.]

The estimator for the status quo average outcome is just the average sample outcome, $\hat{\beta}_{sq}=\sum_i Y_i/N$. This is efficient and inference is entirely standard. For the other estimators we first need to estimate the regression function $g(w,x,z)$. We estimate $g(w,x,z)$ using nonparametric methods. We use kernel methods, although series estimators could also be used. For a kernel $K(u)$, with $u \in \mathbb{R}^{K+2}$, bandwidth b, and with $V_i = (W_i, X_i, Z_i')'$ and $v = (w, x, z')'$, we have

$$\hat{g}(w,x,z) = \frac{\sum_{i=1}^{N} Y_i \cdot K((v - V_i)/b)}{\sum_{i=1}^{N} K((v - V_i)/b)}.$$

In the sequel we use the notation $K_b(v) = \frac{1}{b^{K+2}} K\left(\frac{v}{b}\right)$.

Each of our estimators also require plug-in estimates of either $F_X(x)$ or $F_W(w)$ or both. For these objects we use the empirical CDFs

$$\hat{F}_X(x) = \frac{1}{N}\sum_{i=1}^{N} 1\{X_i \leq x\}, \qquad \hat{F}_W(w) = \frac{1}{N}\sum_{i=1}^{N} 1\{W_i \leq w\}.$$

For the quantile function $F_W^{-1}(q)$, required for the perfect positive and negative matching cases, we use the inverse of the empirical distribution function of W, i.e.,

$$\hat{F}_W^{-1}(q) = \inf_{w \in \mathbb{W}} 1\{\hat{F}_W(w) \geq q\}.$$

We then estimate β^{pam} and β^{nam} by the analog estimators

$$\widehat{\beta}^{\text{pam}} = \frac{1}{N}\sum_{i=1}^{N} \hat{g}\left(\hat{F}_W^{-1}(\hat{F}_X(X_i)), X_i, Z_i\right),$$

and

$$\widehat{\beta}^{\text{nam}} = \frac{1}{N}\sum_{i=1}^{N} \hat{g}\left(\hat{F}_W^{-1}(1 - \hat{F}_X(X_i)), X_i, Z_i\right).$$

For the purposes of estimation, the correlated matching allocations are redefined using a truncated bivariate normal cupola. The truncation ensures that the denominator in the weights of the correlated matching ARE are bounded from 0, so that we do not require trimming. The bivariate standard normal PDF is

$$\phi(x_1, x_2; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}(x_1^2 - 2\rho x_1 x_2 + x_2^2)},$$

with a corresponding joint CDF denoted by $\Phi(x_1, x_2; \rho)$. Observe that

$$\Pr(-c < x_1 \leq c, -c < x_2 \leq c) = \Phi(c, c; \rho) - \Phi(c, -c; \rho) - [\Phi(-c, c; \rho) - \Phi(-c, -c; \rho)],$$

so that the truncated standard bivariate normal PDF is given by

$$\phi_c(x_1, x_2; \rho) = \frac{\phi(x_1, x_2; \rho)}{\Phi(c, c; \rho) - \Phi(c, -c; \rho) - [\Phi(-c, c; \rho) - \Phi(-c, -c; \rho)]}$$

with $-c < x_1, x_2 \leq c$. Denote the truncated bivariate CDF by Φ_c.

The truncated normal bivariate CDF gives a comprehensive cupola, because the corresponding joint CDF

$$H_{W,X}(w,x) = \Phi_c\left(\Phi_c^{-1}(F_W(w)), \Phi_c^{-1}(F_X(x)); \rho\right)$$

has marginal CDFs equal to $H_{W,X}(w,\infty) = F_W(w)$ and $H_{W,X}(\infty, x) = F_X(x)$, it reaches the upper and lower Fréchet bounds on the joint CDF for $\rho = 1$ and $\rho = -1$, respectively, and it has independent W, X as a special case for $\rho = 0$.

To obtain an estimate of $\beta^{\text{cm}}(\rho, \tau)$ we note that joint PDF associated with $H_{W,X}(w, x)$ equals

$$h_{W,X}(w,x) = \phi_c\left(\Phi_c^{-1}(F_W(w)), \Phi_c^{-1}(F_X(x)); \rho\right) \frac{f_W(w) f_X(x)}{\phi_c\left(\Phi_c^{-1}(F_W(w))\right) \phi_c\left(\Phi_c^{-1}(F_X(x))\right)},$$

and hence that $\beta^{\text{cm}}(\rho, 0)$, redefined in terms of the truncated normal, is given by

$$\beta^{\text{cm}}(\rho, 0) = \int_{x,z} \int_w g(w,x,z) \frac{\phi_c\left(\Phi_c^{-1}(F_W(w)), \Phi_c^{-1}(F_X(x)); \rho\right)}{\phi_c\left(\Phi_c^{-1}(F_W(w))\right) \phi_c\left(\Phi_c^{-1}(F_X(x))\right)} f_W(w) f_{X,Z}(x,z) \mathrm{d}w \mathrm{d}x \mathrm{d}z.$$

Replacing the integrals with sums over the empirical distribution we get the analog estimator

$$\widehat{\beta}^{\text{cm}}(\rho, 0) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \hat{g}(W_i, X_j, Z_j) \frac{\phi_c\left(\Phi_c^{-1}(\hat{F}_W(W_i)), \Phi_c^{-1}(\hat{F}_X(X_j)); \rho\right)}{\phi_c\left(\Phi_c^{-1}(\hat{F}_W(W_i))\right) \phi_c\left(\Phi_c^{-1}(\hat{F}_X(X_j))\right)}.$$

Observe that if $\rho = 0$ (independent matching) the ratio of densities on the right hand side is equal to 1.

For $\tau > 0$, the $\beta^{\text{cm}}(\rho, \tau)$ estimand is a convex combination of average output under the status quo and a correlated matching allocation. The corresponding sample analog is

$$\widehat{\beta}^{\text{cm}}(\rho, \tau) = \tau \cdot \widehat{\beta}^{\text{sq}} + (1 - \tau) \cdot \widehat{\beta}^{\text{cm}}(\rho, 0).$$

This estimator is linear in the nonparametric regression estimator $\hat{g}$ and nonlinear in the empirical CDFs of X and W. This structure simplifies the asymptotic analysis.

4.2 Asymptotic properties of the correlated matching estimator

A useful and insightful representation of $\beta^{\text{cm}}(\rho, 0)$ is as an average of partial means (c.f., Newey 1994). This representation provides intuition both about the structure of the estimand as well as its large sample properties. Fixing W at $W = w$ but averaging over the joint distribution of X and Z we get the partial mean:

$$\eta(w) = \mathbb{E}_{X,Z}\left[g(w, X, Z) \times d(w, X)\right], \tag{4.5}$$

where

$$d(w,x) = \frac{\phi_c(\Phi_c^{-1}(F_W(w)), \Phi_c^{-1}(F_X(x)); \rho)}{\phi_c(\Phi_c^{-1}(F_W(w)))\phi_c(\Phi_c^{-1}(F_X(x)))}. \tag{4.6}$$

Observe that (4.5) is a weighted averaged of the production function over the joint distribution of X and Z holding the value of the input to be reallocated W fixed at $W = w$. The weight function $d(w, X)$ depends upon the truncated normal cupola. In particular, the weights give greater emphasis to realizations of $g(w, X, Z)$ that are associated with values of X that will be assigned a value of W close to w as part of the correlated matching reallocation. Thus (4.5) equals the average post-reallocation output for those firms being assigned $W = w$. To give a concrete example (4.5) is the post-reallocation expected achievement of those classrooms that will be assigned a teacher of quality $W = w$.

Equation (4.5) also highlights the value of using the truncated normal copula. Doing so ensures that the denominators of the copula 'weights' in (4.5) are bounded from zero. The copula weights thus play the role similar to fixed trimming weights used by Newey (1994).

If we average these partial means over the marginal distribution of W we get $\beta^{\text{cm}}(\rho, 0)$, since

$$\beta^{\text{cm}}(\rho, 0) = \mathbb{E}_W \left[\eta\left(W\right)\right],$$

yielding average output under the correlated matching reallocation.

From the above discussion it is clear that our correlated matching estimator can be viewed as a semiparametric two-step method-of-moments estimator with a moment function of

$$m(Y, W, \beta^{\text{cm}}(\rho, \tau), \eta\left(W\right)) = \tau Y + (1 - \tau)\,\eta\left(W\right) - \beta^{\text{cm}}(\rho, \tau).$$

Our estimator, $\widehat{\beta}^{\text{cm}}(\rho, \tau)$, is the feasible GMM estimator based upon the above moment function after replacing the partial mean (4.5) with a consistent estimate. While the above representation is less useful for deriving the asymptotic properties of $\widehat{\beta}^{\text{cm}}(\rho, \tau)$ it does provide some insight as to why we are able to achievement parametric rates of convergence.

To derive the asymptotic distribution of $\widehat{\beta}^{\text{cm}}(\rho, \tau)$ we show that $\widehat{\beta}^{\text{cm}}(\rho, 0)$ is asymptotically linear (this is sufficient since the sample mean of Y is clearly a linear estimator and $\widehat{\beta}^{\text{cm}}(\rho, \tau)$ is a linear combination of $\overline{Y}$ and $\widehat{\beta}^{\text{cm}}(\rho, 0)$). This result is obtained in two steps. First, we express $\widehat{\beta}^{\text{cm}}(\rho, \tau) - \beta^{\text{cm}}(\rho, \tau)$ as a sum of U-statistics. Second, we determine the projections of these U-statistics to derive the asymptotically linear representation.

For the first step we follow the approach in Newey (1994). In particular, we define the functions

$$k_{1,ijk}(\zeta) = \frac{\hat{h}_1(W_j, X_k, Z_k) + \zeta Y_i K_b(W_j - W_i, X_k - X_i, Z_k - Z_i)}{\hat{h}_2(W_j, X_k, Z_k) + \zeta K_b(W_j - W_i, X_k - X_i, Z_k - Z_i)} \tag{4.7}$$

and

$$k_{2,ijk}(\zeta) = \frac{\phi_c(\Phi_c^{-1}(\widehat{F}_W(W_j) + \zeta I(W_j \leq W_i)), \Phi_c^{-1}(\widehat{F}_X(X_k) + \zeta I(X_k \leq X_i)); \rho)}{\phi_c(\Phi_c^{-1}(\widehat{F}_W(W_j) + \zeta I(W_j \leq W_i)))\phi_c(\Phi_c^{-1}(\widehat{F}_X(X_k) + \zeta I(X_k \leq X_i)))} \tag{4.8}$$

where $\hat{h}_1(w, x, z)$ and $\hat{h}_2(w, x, z)$ respectively equal

$$\hat{h}_1(w, x, z) = \frac{1}{N}\sum_{i=1}^{N} y_i K_b(w - W_i, x - X_i, z - Z_i) \tag{4.9}$$

and

$$\hat{h}_2(w, x, z) = \frac{1}{N}\sum_{i=1}^{N} K_b(w - W_i, x - X_i, z - Z_i), \tag{4.10}$$

such that the kernel estimate of the production function is equal to their ratio, i.e., $\hat{g}(w, x, z) = \hat{h}_1(w, x, z)/\hat{h}_2(w, x, z)$.

Taking product of (4.7) and (4.8) and summing over all j and k we have

$$\frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} k_{1,ijk}(\zeta)k_{2,ijk}(\zeta). \tag{4.11}$$

Observe that

$$\frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} k_{1,ijk}(\zeta)k_{2,ijk}(\zeta)\bigg|_{\zeta=0} = \frac{1}{N}\sum_{j=1}^{N}\widehat{\eta}\left(W_j\right) = \widehat{\beta}^{\text{cm}}(\rho,0).$$

Differentiating (4.11) with respect to ζ and evaluating at $\zeta = 0$ therefore provides an estimate of the influence of the i^{th} observation in $\widehat{\gamma}$ – the vector of estimated partial means which are averaged over in (4.11) – on $\widehat{\beta}^{\text{cm}}(\rho,0)$. We have

$$\frac{\partial k_{1,ijk}}{\partial\zeta}(\zeta)\bigg|_{\zeta=0} = \frac{K_b(W_j-W_i,X_k-X_i,Z_k-Z_i)}{\hat{h}_2(W_j,X_k,Z_k)}\times\left(Y_i-\frac{\hat{h}_1(W_j,X_k,Z_k)}{\hat{h}_2(W_j,X_k,Z_k)}\right),$$

and using the derivatives of the cupola in Appendix A.2 we also have

$$\frac{\partial k_{2,ijk}}{\partial\zeta}(\zeta)\bigg|_{\zeta=0} = \hat{e}_W(W_j,X_k)I(W_i\leq W_j)+\hat{e}_X(W_j,X_k)I(X_i\leq X_k),$$

where

$$\begin{aligned}\hat{e}_W(W_j,X_k) &= \frac{\rho\phi_c(\Phi_c^{-1}(\widehat{F}_W(W_j)),\Phi_c^{-1}(\widehat{F}_X(X_k));\rho)}{(1-\rho^2)\phi_c(\Phi_c^{-1}(\widehat{F}_W(W_j)))^2\phi_c(\Phi_c^{-1}(\widehat{F}_X(X_k)))}\times \\ &\quad\left[\Phi_c^{-1}(\widehat{F}_X(X_k))-\rho\Phi_c^{-1}(\widehat{F}_W(W_j))\right]\end{aligned} \tag{4.12}$$

$$\begin{aligned}\hat{e}_X(W_j,X_k) &= \frac{\rho\phi_c(\Phi_c^{-1}(\widehat{F}_W(W_j)),\Phi_c^{-1}(\widehat{F}_X(X_k));\rho)}{(1-\rho^2)\phi_c(\Phi_c^{-1}(\widehat{F}_W(W_j)))\phi_c(\Phi_c^{-1}(\widehat{F}_X(X_k)))^2}\times \\ &\quad\left[\Phi_c^{-1}(\widehat{F}_W(W_j))-\rho\Phi_c^{-1}(\widehat{F}_X(X_k))\right].\end{aligned} \tag{4.13}$$

Our estimated influence function is therefore given by

$$\widehat{\psi}_i = m(Y_i,W_i,\widehat{\beta}^{\text{cm}}(\rho,0),\widehat{\gamma}\left(W_i\right))+\frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\left[\frac{\partial k_{1,ijk}}{\partial\zeta}(\zeta)\bigg|_{\zeta=0}\times k_{2,ijk}(\zeta)+k_{1,ijk}\times\frac{\partial k_{2,ijk}}{\partial\zeta}(\zeta)\bigg|_{\zeta=0}\right], \tag{4.14}$$

where the second and third terms are estimates of the first-order effect of the i^{th} observation's effect on sampling variation in moment function operating through the estimated regression function $g\left(w,x,z\right)$ and CDFs of W and X respectively.

Replacing estimates with population values in (4.14), summing over $i=1,\ldots,N$ and letting

$b_N \downarrow 0$ as $N \to \infty$ we obtain the following asymptotically equivalent expression for $\widehat{\beta}^{\text{cm}}(\rho, 0)$

$$\widehat{\beta}^{\text{cm}}(\rho,0) - \beta^{\text{cm}}(\rho,0) = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} g(W_j, X_k, Z_k) d(W_j, X_k) - \beta^{\text{cm}}(\rho,0)$$

$$+ \frac{1}{N}\sum_{i=1}^{N} \frac{f_W(W_i) f_{XZ}(X_i, Z_i)}{f_{W,X,Z}(W_i, X_i, Z_i)} (Y_i - g(W_i, X_i, Z_i)) d(W_i, X_i)$$

$$+ \frac{1}{N^3}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} e_W(W_j, X_k)(I(W_i \leq W_j) - F_W(W_j)) g(W_j, X_k, Z_k)$$

$$+ \frac{1}{N^3}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} e_X(W_j, X_k)(I(X_i \leq X_k) - F_X(X_k)) g(W_j, X_k, Z_k),$$

with $d(W_j, X_k)$ as defined in (4.6) above and $e_W(W_j, X_k)$ and $e_X(W_j, X_k)$ as given by (4.12) and (4.13) but with population values replacing estimates.

The first, third and fourth lines of the above expression are (two-sample) U-statistics. The final asymptotically linear expression is obtained by projection of these statistics

$$\widehat{\beta}^{\text{cm}}(\rho,0) - \beta^{\text{cm}}(\rho,0) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[g(W, X_i, Z_i) d(W, X_i)\right] - \beta_{cs}$$

$$+ \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[g(W_i, X, Z) d(W_i, X)\right] - \beta_{cs}$$

$$+ \frac{1}{N}\sum_{i=1}^{N} \frac{f_W(W_i) f_{XZ}(X_i, Z_i)}{f_{WXZ}(W_i, X_i, Z_i)} (Y_i - g(W_i, X_i, Z_i)) d(W_i, X_i)$$

$$+ \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[e_W(W, X) g(X, W, Z)(I(W_i \leq W) - F_W(W))\right]$$

$$+ \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[e_W(W, X) g(X, W, Z)(I(X_i \leq X) - F_X(X))\right],$$

where the expectations are over the product of the marginal distributions of W and X, Z. Note that the term that accounts for the estimation of the regression function is uncorrelated with all other terms, which simplifies the expression for the asymptotic variance.

In order to state a formal result we need the following assumptions

Assumption 4.1 *Let $v = (w\ x\ z')'$. The function $K(v)$ is bounded on a bounded set $\mathcal{V}$ and $K(v) = 0$ for $v \in \mathcal{V}^c$. Also K is a kernel of order S*

$$\int_{\mathcal{V}} K(v)dv = 1 \qquad \int_{\mathcal{V}} v^s K(v)dv = 0$$

for $s = 1, \ldots, S$ with $v^s = \prod_{s_1 \geq 0, \ldots, s_{K+2} \geq 0, s_1 + \cdots s_{K+2} = s} v_1^{s_1} \cdots v_{K+2}^{s_{K+2}}$

Assumption 4.2 *The joint density $f_{W,X,Z}(w, x, z)$ has compact support $\mathcal{W} \times \mathcal{X} \times \mathcal{Z}$ and the density is bounded from 0 and ∞ on this support.*

Assumption 4.3 *The function* $g(w,x,z) = \mathbb{E}[Y|W=w, X=x, Z=z]$ *that is defined on* $\mathcal{W}\times\mathcal{X}\times\mathcal{Z}$ *can be extended to* $\Re^{K+2}$ *such that it is* S *times continuously differentiable and the* S*-th derivative is bounded on* $\Re^{K+2}$*.*

Assumption 4.4 $\mathbb{E}[Y^2|W,X,Z]$ *is bounded on* $\mathcal{W}\times\mathcal{X}\times\mathcal{Z}$*.*

Assumption 4.5 *The bandwidth sequence is such that as* $N\to\infty$

$$\frac{N^{\frac{1}{4}}}{\sqrt{\ln N}} b_N^{\frac{K}{2}+1} \to \infty, \qquad \sqrt{N} b_N^{S(K+2)} \to 0.$$

Theorem 4.1 *If assumptions 4.1-4.5 hold, then*

$$\widehat{\beta}^{\text{cm}}(\rho,\tau) \xrightarrow{p} \beta^{\text{cm}}(\rho,\tau)$$

and

$$\sqrt{N}(\widehat{\beta}^{\text{cm}}(\rho,\tau) - \beta^{\text{cm}}(\rho,\tau)) = \tau\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(y_i - \beta^{\text{sq}}) + (1-\tau)\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi_i + o_p(1)$$

with

$$\begin{aligned}\psi_i = {} & \mathbb{E}\left[g(W,X_i,Z_i)d(W,X_i)\right] - \beta^{\text{cm}}(\rho,0) + \mathbb{E}\left[g(W_i,X,Z)d(W_i,X)\right] - \beta^{\text{cm}}(\rho,0) \\ & + \frac{f_W(W_i)f_{XZ}(X_i,Z_i)}{f_{WXZ}(W_i,X_i,Z_i)}(Y_i - g(W_i,X_i,Z_i))d(W_i,X_i) \\ & + \mathbb{E}\left[e_W(W,X)g(X,W,Z)(I(W_i\le W) - F_W(W))\right] \\ & + \mathbb{E}\left[e_W(W,X)g(W,X,Z)(I(X_i\le X) - F_X(X))\right].\end{aligned} \tag{4.15}$$

and $d(w,x)$*,* $e_X(w,x)$ *and* $e_W(w,x)$ *as defined above.*

Proof. See Appendix A.1 ■

A consistent estimate of the asymptotic variance of $\widehat{\beta}^{\text{cm}}$ is given by $\frac{1}{N}\sum_{i=1}^{N}(\tau(Y_i - \beta^{\text{sq}}) + (1-\tau)\widehat{\psi}_i)^2$, where $\widehat{\psi}_i$ is as defined in (4.14) above.

4.3 Asymptotic properties of the perfect assortative matching estimator

In this subsection we discuss the large sample properties of $\hat{\beta}^{\text{pam}}$. The rate of convergence of $\hat{\beta}^{\text{pam}}$ to β^{pam} is slower than the regular parametric rate. This is because we estimate a nonparametric regression function with more arguments than we average over in the second stage. We will show that

$$\hat{\beta}^{\text{pam}} - \beta^{\text{pam}} = O_p\left(N^{-1/2}b_N^{-1/2}\right).$$

However, in order to improve the performance of confidence intervals we will take into account additional terms in the asymptotic expansion beyond the leading $O_p(N^{-1/2}b_N^{-1/2})$ term. Specifically, we take into account the $O_p(N^{-1/2})$ terms.

Formally, we will show that

$$\hat{\beta}^{\text{pam}} = \beta^{\text{pam}} + \beta^{\text{pam}} + b_N^{-1/2}\hat{\mu}_Y + \hat{\mu}_W + \hat{\mu}_X + \hat{\mu}_g + o_p\left(N^{-1/2}\right),$$

with $\hat{\mu}_Y$, $\hat{\mu}_W$, $\hat{\mu}_X$, and $\hat{\mu}_g$ sample averages satisfying a joint central limit theorem:

$$\sqrt{N} \cdot \begin{pmatrix} \hat{\mu}_Y \\ \hat{\mu}_W \\ \hat{\mu}_X \\ \hat{\mu}_g \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \Omega \right).$$

We propose a consistent estimator $\hat{\Omega}$ for Ω. As the estimator for the variance of $N^{1/2} b_N^{1/2} (\hat{\beta}^{\text{pam}} - \beta^{\text{pam}}$ we then propose to use

$$\hat{\sigma}_\beta^2 = \begin{pmatrix} 1 \\ b_N^{1/2} \\ b_N^{1/2} \\ b_N^{1/2} \end{pmatrix}' \hat{\Omega} \begin{pmatrix} 1 \\ b_N^{1/2} \\ b_N^{1/2} \\ b_N^{1/2} \end{pmatrix}.$$

As $N \to \infty$, this converges to

$$\text{plim}_{N\to\infty} \hat{\sigma}_\beta^2 = \Omega_{11},$$

because $b_N \to 0$. Hence we could simply estimate the asymptotic variance as

$$\tilde{\sigma}_\beta^2 = \hat{\Omega}_{11}.$$

Nevertheless, it is likely that taking into account the variation in $\hat{\mu}_W$, $\hat{\mu}_X$ and $\hat{\mu}_g$ will improve the finite sample properties of the confidence intervals.

We are interested in the asymptotic distribution of

$$\hat{\beta}^{\text{pam}} = \frac{1}{N} \sum_{i=1}^{N} \hat{g}\left(\hat{F}_W^{-1}(\hat{F}_X(X_i)), X_i \right).$$

Assumption 4.6 *There are positive integers Δ and s such that $\mathcal{K}(u)$ is differentiable of order Δ, the derivatives of order Δ are Lipschitz, $\mathcal{K}(u)$ is zero outside a bounded set, $\int \mathcal{K}(u)du = 1$, and for all $j < s$, $\int \mathcal{K}(u)[\otimes_{l=1}^{j}]du = 0$.*

quote from newey: "The last condition requires that the kernel be a higher order (bias reducing) kernel of order s."

Assumption 4.7 (PROBABLY SUPERFLUOUS GIVEN OTHER ASSUMPTIONS) *There is a non-negative integer d and an extension of $g(x)$ to all of $\mathbb{R}^k$ that is continuously differentiable to order d on $\mathbb{R}^k$.*

Assumption 4.8 (PROBABLY SUPERFLUOUS GIVEN OTHER ASSUMPTIONS) *For $p \geq 4$, $\mathbb{E}[|Y|^p] \leq \infty$, $\mathbb{E}[|Y|^p|X = x]f_(x)$ is bounded, $\mathbb{E}[\|m(z, \beta_0, g)\|^2] < \infty$.*

Assumption 4.9 (SMOOTHNESS OF $g(w, x)$)
$g(w, x)$ is twice continuously differentiable with respect to w on $\mathbb{W} \times \mathbb{X}$.

Assumption 4.10 (Distribution of Data)
(i) The support of W is $\mathbb{W}$, a compact subset of $\mathbb{R}$,
(ii) the support of X is $\mathbb{X}$, a compact subset of $\mathbb{R}$,
(iii) the joint distribution of W and X is bounded and bounded away from zero on $\mathbb{W} \times \mathbb{X}$,
(iv) the conditional expectations $\mu_4(v) = \mathbb{E}[Y^4|W = w, X = x]$ is bounded.

Assumption 4.11 *The bandwidth b_N satisfies $b_N \to 0$, $N^{-1}b_N^{-2k_1} \to 0$ $N^{2/p-1}b_N^{-k}\ln(N) \to 0$, ($p$ is moment of Y that exists)*

(Whitney assumes $(Nb_N^{k_1})^{-1}$ converges to zero, but he uses a complicated condition on the fourth moment of Y given V in Assumption 5.1 that may be more restrictive. It is a little unclear to me).

First we decompose $\hat{\beta}^{\text{pam}} - \beta^{\text{pam}}$ into four parts plus a lower order remainder term. The first part corresponds to the uncertainty in $\hat{g}(\cdot, \cdot)$, the second corresponds to the uncertainty in $\hat{F}_W^{-1}(\cdot)$, the third part corresponds to the uncertainty in $\hat{F}_X(\cdot)$, and the final part corresponds to the difference between the average of $g\left(F_W^{-1}(\hat{F}_X(X_i)), X_i\right)$ and its expectation.

Lemma 4.1

$$\hat{\beta}^{\text{pam}} - \beta^{\text{pam}} = \frac{1}{N}\sum_{i=1}^{N} \hat{g}\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) - \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) \tag{4.16}$$

$$+\frac{1}{N}\sum_{i=1}^{N} g\left(\hat{F}_W^{-1}\left(F_X(X_i)\right), X_i\right) - \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) \tag{4.17}$$

$$+\frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(\hat{F}_X(X_i)\right), X_i\right) - \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) \tag{4.18}$$

$$+\frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) - \mathbb{E}\left[g\left(F_W^{-1}\left(F_X(X)\right), X\right)\right] + o_p\left(N^{-1/2}\right). \tag{4.19}$$

Define

$$g_w(w,x) = \frac{\partial g}{\partial w}(w,x), \qquad \text{and} \quad g_{ww}(w,x) = \frac{\partial^2 g}{\partial w^2}(w,x),$$

$$q_{WX}(w,x) = \frac{g_w(F_W^{-1}(F_X(x)),x)}{f_W(F_W^{-1}(F_X(x)))} \cdot \left(1\{F_W(w) \leq F_X(x)\} - F_X(x)\right),$$

$$\hat{\mu}_{WX} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} q_{WX}(W_i, X_j),$$

$$q_W(w) = \mathbb{E}[q_{WX}(w,X)], \qquad \text{and} \quad q_X(x) = \mathbb{E}[q_{WX}(W,x)],$$

$$\hat{\mu}_W = \frac{1}{N}\sum_{i=1}^{N} q_W(W_i),$$

$$r_{X_1X_2}(x_1,x_2) = \frac{g_w(F_W^{-1}(F_X(x_2)),x_2)}{f_W(F_W^{-1}(F_X(x_2)))} \cdot (1\{x_1 \le x_2\} - F_X(x_2)),$$

$$\hat{\mu}_{X_1X_2} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} r_{X_1X_2}(X_i,X_j),$$

$$r_X(x) = \mathbb{E}[r_{X_1X_2}(x,X)], \qquad \text{and} \quad r_Z(x) = \mathbb{E}[r_{X_1X_2}(X,x)],$$

$$\hat{\mu}_X = \frac{1}{N}\sum_{i=1}^{N} r_X(X_i).$$

Next we give asymptotically linear representations for the first three components of the difference $\hat{\beta}^{\text{pam}} - \beta^{\text{pam}}$ (the fourth is already a sample average). Define

$$\psi_{Ni} = \sigma_N^{k_1/2} \cdot \int \left(Y_i - g(F_W^{-1}(F_X(u)),u)\right) \cdot K_{\sigma_N}(F_W^{-1}(F_X(u)) - W_i, u - X_i)du,$$

Lemma 4.2 (Asymptotic Linearity)
(i)

$$\frac{1}{N}\sum_{i=1}^{N} g\left(\hat{F}_W^{-1}\left(F_X(X_i)\right), X_i\right) - \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) = \frac{1}{N}\sum_{i=1}^{N} q_W(W_i) + o_p\left(N^{-1/2}\right)$$

(ii)

$$\frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(\hat{F}_X(X_i)\right), X_i\right) - \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) = \frac{1}{N}\sum_{i=1}^{N} r_X(X_i) + o_p\left(N^{-1/2}\right)$$

(iii)

$$\frac{\sigma^{k_1/2}}{N} \cdot \sum_{i=1}^{N}\Big(\hat{g}\left(F_W^{-1}F_X(X_i)), X_i\right) - g\left(F_W^{-1}F_X(X_i)), X_i\right)\Big) = \frac{1}{N} \cdot \sum_{i=1}^{N} \psi_{\sigma i} + o_p\left(N^{-1/2}\right).$$

Next, we give asymptotic normality results for the asymptotic linear representations. Define:

$$\Omega_{33} = \int\int\left[\int K\left(\frac{\partial F_W^{-1}(F_X(v_2))}{\partial v_2}u_2 + u_1, u_2\right)du_2\right]^2 du_1 \sigma^2(F_W^{-1}(F_X(v_2)),v_2)) f_{W,X}(F_W^{-1}(F_X(v_2)),v_2)dv_2$$

Lemma 4.3

$$\frac{1}{N^{1/2}} \cdot \sum_{i=1}^{N} \begin{pmatrix} q_W(W_i) \\ r_X(X_i) \\ (\psi_{\sigma i} - \mathbb{E}[\psi_{\sigma i}]) \\ g(F_W^{-1}F_X(X_i)) - \mathbb{E}\left[g(F_W^{-1}F_X(X))\right] \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \Omega\right).$$

All remaining results are collected in Appendix A.3.

5 A test for 'local' complementarity of w and x

A potential problem with the $\beta(\rho,\tau)$ family of estimands is that the support requirements for their precise estimation may be difficult to satisfy in practice, particularly for allocations 'distant' from the status quo. For this reason a measure of local (to the status quo) complementarity between W and X would be valuable. To this end we next characterize the mean effect associated with a 'small' increase toward either positive or negative assortative matching. The resulting estimand forms the basis of a simple test for local efficiency of the status quo allocation.

We implement our local reallocation as follows: for $\lambda \in [-1,1]$, let $W_\lambda = \lambda \cdot X + (\sqrt{1-\lambda^2}) \cdot W$ be a random variable indexed by λ. The average output associated with positive assortative matching on W_λ is given by

$$\beta^{\mathrm{lr}}(\lambda) = \mathbb{E}[g(F_{W|Z}^{-1}(F_{W_\lambda|Z}(W_\lambda|Z)|Z), X, Z)]. \tag{5.20}$$

For $\lambda = 0$ and $\lambda = 1$ we have $W_\lambda = W$ and $W_\lambda = X$ respectively and hence $\beta^{\mathrm{lr}}(0) = \beta^{\mathrm{sq}}$ and $\beta^{\mathrm{lr}}(1) = \beta^{\mathrm{pam}}$. Perfect negative assortative matching is also nested in this framework since

$$\Pr(-X \leq -x|Z) = 1 - F_{X|Z}(x|Z),$$

and hence for $\lambda = -1$ we have $\beta^{\mathrm{lr}}(-1) = \beta^{\mathrm{nam}}$. Values of λ close to zero induce reallocations of W that are 'local' to the status quo, with $\lambda > 0$ and $\lambda < 0$ generating shifts toward positive and negative assortative matching respectively.

The *sign* of the effect on average outcomes associated with a small step away from the status quo and toward positive assortative matching is given by the sign of

$$\gamma = \frac{\partial \beta^{\mathrm{lr}}}{\partial \lambda}(0), \tag{5.21}$$

while that associated with a small step toward negative assortative matching is given by the sign of $-\gamma$.

Equation (5.21) has two alternative representations which are given in the following Lemma.

Lemma 5.1 *$\gamma = \partial \beta^{lr}(0)/\partial \lambda$ has equivalent representations of*

$$\gamma = \mathbb{E}\left[\frac{\partial g}{\partial w}(W,X,Z) \cdot (X - m(W,X))\right], \tag{5.22}$$

where $m(w,z) = \mathbb{E}[X|W=w, Z=z]$ and, if the support of X is bounded (i.e., $a \leq X \leq b$), of

$$\gamma = \mathbb{E}\left[Var(X|W,Z) \cdot \mathbb{E}_{X|W,Z}\left[\omega(V)\frac{\partial^2 g}{\partial w \partial x}(W,X,Z)|W,Z\right]\right], \tag{5.23}$$

where $V = (W,X,Z')'$ as above and

$$\omega(W,t,Z) = \frac{1}{dF_{X|W,Z}(t|W,Z)} \frac{\mathbb{E}_{X|W,Z}[X - m(W,X)|W,Z,X \geq t]\left(1 - F_{X|W,Z}(t|W,Z)\right)}{\int_{r=a}^{r=b} \mathbb{E}_{X|W,Z}[X - m(W,X)|W,Z,X \geq r]\left(1 - F_{X|W,Z}(r|W,Z)\right) dr}$$

are weights with a population mean of 1 (i.e., $\mathbb{E}_{X|W,Z}[\omega(V)|W,Z] = 1$) and which emphasize values of $\frac{\partial^2 g}{\partial w \partial x}(W,X,Z)$ where X is near its conditional mean, $m(W,X)$.

Proof. See Appendix A.4. ■

Representation (5.22), as we demonstrate below, suggests a straightforward method-of-moments approach to estimating γ_0. Representation (5.23) is valuable for interpretation. Equation (5.23) demonstrates that a test of $H_0 : \gamma = 0$ is a test of the the null of no complementarity or substitutability between W and X. If $\gamma > 0$, then in the 'vicinity of the status quo' W and X are complements; if $\gamma < 0$ they are substitutes. The precise meaning of the 'vicinity of the status quo' is implicit in the form of the weight function $\omega(V)$.

Deviations of γ from zero imply that the status quo allocation does not maximize average outcomes. For $\gamma > 0$ a shift toward positive assortative matching will raise average outcomes, while for $\gamma < 0$ a shift toward negative assortative matching will do so. Lemma 5.1 therefore provides the basis of a test for whether the status quo allocation is locally efficient.

Estimation of γ proceeds in two-steps. First we estimate $g(w, x, z)$ and $m(w, z)$ using kernel methods as in Section 4. In the second step we estimate γ by method-of-moments using the sample analog of the moment condition

$$\mathbb{E}[m(Y, V, \gamma_0, g, m)] = \mathbb{E}[\nabla_w g(W, X, Z)(X - m(W, Z)) - \gamma_0] = 0,$$

where $g(W, X, Z)$ and $m(W, Z)$ are replaced with the first step estimates, i.e.,

$$\widehat{\gamma} = \frac{1}{N}\sum_{i=1}^{N} \nabla_w \widehat{g}(W_i, X_i, Z_i) \times (X_i - \widehat{m}(W_i, Z_i)). \tag{5.24}$$

Note that we compute $\nabla_w \widehat{g}(w, x, z)$ by analytically differentiating $\widehat{g}(w, x, z)$ with respect to w.

The asymptotic properties of $\widehat{\gamma}$ are derived analogously to those of $\widehat{\beta}^{\text{cm}}$ and are summarized by Theorem 5.1.

Theorem 5.1 *Under conditions 4.1 to 4.5 $\widehat{\gamma}$ is $\sqrt{N}$ consistent and asymptotically normal, i.e.,*

$$\widehat{\gamma} \xrightarrow{p} \gamma$$

and

$$\sqrt{N}(\widehat{\gamma} - \gamma_0) \xrightarrow{D} \mathcal{N}(0, \Lambda_0), \quad \Lambda_0 = Var(\psi)$$

where $\psi = m(Y, V, \gamma_0, g, m) + \delta(Y, V)$ with

$$\begin{aligned}\delta(Y, V) = &-\frac{1}{f_{W,X,Z}(W, X, Z)}\frac{\partial f_{W,X,Z}(W, X, Z)}{\partial W}(Y - g(W, X, Z))(X - m(W, Z)) \\ &- \frac{\partial m(W, Z)}{\partial W}(Y - g(W, X, Z)) \\ &\qquad - \mathbb{E}\left[\frac{\partial g(W, X, Z)}{\partial W}|W = w, Z = z\right](X - m(W, Z)).\end{aligned}$$

Proof. See Appendix A.5 ■

Theorem 5.1 follows from the fact that $\widehat{\gamma}$ admits an asymptotically linear representation of

$$\widehat{\gamma} = \gamma_0 + \frac{1}{N}\sum_{i=1}^{N} \{m(Y_i, V_i, \gamma, g, m) + \delta(Y_i, V_i)\} + o_p(1/\sqrt{N}). \tag{5.25}$$

If $\widehat{g}(w, x, z)$ and $\widehat{m}(x, z)$ are replaced by their population values in (**??**) the final three terms in (5.25) drop out; these terms therefore represent the effect on the moment function of replacing $g(w, x, z)$ and $m(x, z)$ with their nonparametric first step estimates. The first two capture the effect of sampling error in $\nabla_w \widehat{g}(w, x, z)$ on the large sample behavior of $m(Y, V, \gamma, g, m)$, while the final one captures the effect of sampling error in $\widehat{m}(x, z)$.

6 Estimation and inference with Discretely-valued inputs

In Section 4 we assumed that both inputs (W, X) are continuous. Fortunately we can use the same basic approach for discretely-valued inputs. Only for $\rho = 1$ or $\rho = -1$ (perfect sorting) are there some complications.

We first consider $-1 < \rho < 1$. The general formula for a bivariate discrete density if W takes the values $w_1, \ldots, w_K$ with probabilities $p_1, \ldots, p_K$ and X takes the values $x_1, \ldots, x_L$ with probabilities $q_1, \ldots, q_L$ is given by

$$\begin{aligned} h_{W,X}(w_k, x_l) &= \Pr(W = w_k, X = x_l) \\ &= H_{W,X}(w_k, x_l) - H_{W,X}(w_k, x_l-) - H_{W,X}(w_k-, x_l) + H_{W,X}(w_k-, x_l-) \\ &= H_{W,X}(w_k, x_l) - H_{W,X}(w_k, x_{l-1}) - [H_{W,X}(w_{k-1}, x_l) - H_{W,X}(w_{k-1}, x_{l-1})] \end{aligned}$$

Because $F_W(w_k) = \sum_{i=1}^{k} p_i$ and $F_X(x_l) = \sum_{j=1}^{l} q_j$ we have

$$H_{W,X}(w_k, x_l) = \Phi\left(\Phi^{-1}\left(\sum_{i=1}^{k} p_i\right), \Phi^{-1}\left(\sum_{j=1}^{l} q_j\right); \rho\right),$$

and, as can be easily verified,

$$\mathrm{d}\Phi\left(\Phi^{-1}\left(\sum_{i=1}^{k} p_i\right), \Phi^{-1}\left(\sum_{j=1}^{l} q_j\right); 0\right) = (F_W(w_k) - F_W(w_{k-1}))(F_X(x_l) - F_X(x_{l-1})),$$

which is just a product of the marginal probability mass functions $p_k = \mathrm{d}F_W(w_k)$ and $q_l = \mathrm{d}F_W(w_l)$.

The average output given correlated matching is

$$\beta^{\mathrm{cm}}(\rho, 0) = \int_z \sum_{k=1}^{K} \sum_{l=1}^{L} g(w_k, x_l, z) h_{W,X}(w_k, x_l) \mathrm{d}F_{Z|X}(z|x_l)$$

The estimator replaces p_k and q_l by the observed fractions $\hat{p}_k$ and $\hat{q}_l$. If there is no Z or if Z is discrete, $g(w_k, x_l, z)$ is estimated by the sample average of Y for the observations with $W = w_k$, $X = x_l$, and $Z = z$. Inference in this case is entirely standard. Standard errors can be computed using the delta method; Appendix A.6 gives the required formulae.

Now consider the cases where $\rho = 1$ or $\rho = -1$. If $\rho = 1$ the joint distribution of (W, X) is degenerate with a one dimensional support $\{(W, X) | W = F_W^{-1}(F_X(X))\}$. In the discrete case there may be no value of k such that

$$\sum_{i=1}^{k} p_i = \sum_{j=1}^{l} q_j.$$

We therefore need compute a K vector $w(x_l)$ that specifies the fraction of units for each of the K values of W that are assigned to firms with $X = x_l$. That is, $w_k(x_l) = r$ means that a (randomly selected) fraction r of the units with $W = w_k$ is assigned to firms with $X = x_l$.

Because $w = F_W^{-1}(F_X(x))$ means that we match w to x, we can use the following algorithm to determine the K vector $w(x_l)$. First, consider x_1. There are two possibilities $p_1 > q_1$ and $p_1 \leq q_1$. In the first case

$$w_1(x_1) = 1$$

$$w_2(x_1) = \cdots = w_K(x_1) = 0$$

We then set $p_1 = p_1 - q_1$ and consider x_2.

In the second case find k_1 such that

$$\sum_{i=1}^{k_1} p_i \leq q_1 < \sum_{i=1}^{k_1+1} p_i$$

Then

$$w_1(x_1) = \cdots = w_{k_1}(x_1) = 1$$

$$w_{k_1+1}(x_1) = q_1 - \sum_{i=1}^{k_1} p_i$$

$$w_{k_1+2}(x_1) = \cdots = w_K(x_1) = 0$$

Set for $k \geq 2$

$$w_1(x_k) = \cdots = w_{k_1}(x_k) = 0$$

and set

$$p_1 = \sum_{i=1}^{k_1+1} p_i - q_1$$

$1 = k_1 + 1$ and consider x_2.

After the branch specific redefinition of p_1 we compare q_2 to p_1 and branch as above. If we choose the second branch k_1 should be read as k_1 from the previous step plus the new k_1. Repeat this until $l = L$. For $\rho = -1$ we relabel w_k in reverse order and apply the same algorithm.

7 Empirical application: martial sorting and child education

To illustrate our methods in practice we present estimates of AREs from a simple setting. In particular, we consider the effect of parents' education on the education of their child. Kremer (1997) is a related application. He considers the connection between neighborhood and martial sorting in terms of years schooling and inequality in educational attainment among children. Kremer specifies a linear relation between the average level of education of parents and the years of schooling of their children. This implies that the average level of childrens' education is invariant under reallocations of parents.

We use data on 10,272 children from the NLSY to study the relation between the education of parents and the education of their children. Table 1 gives summary statistics.

It should be noted that years of education is not uniformly distributed. In the data 43% of the mothers, 35% of the fathers, and 44% of the children report that they have 12 years of education with further spikes at 16 years of education. Reported years of education vary between 1 and 20. A regression of a child's years of schooling on that of their mother and father (see Table 2) shows that the interaction effect is not significant. The relation is nonlinear however, so that reallocations of parents may affect the average level of child education.

Table 1: Years of education NLSY; $N = 12272$

	Mean	Std. dev.
Ed. child	13.06	2.38
Ed. mother	11.20	2.87
Ed. father	11.20	3.64

Table 2: Regression of education of child on education parents; NLSY, $N = 10272$

	Coefficient	Standard err.
Constant	11.27	.19
Ed. mother	-.041	.036
Ed. father	-.077	.029
Ed. mother^2	.011	.0023
Ed. father^2	.011	.0015
Ed. mother $\times$ Ed. father	.0014	.0029
R^2	.22	

Inspection of the average level of child education cross-classified by parent education shows that a child's educational attainment tends to be high if her mother has a high level of education and her father has a low level of education relative to cases where her mother has a low level of education and her father a high level of education. This asymmetry is not captured by the interaction term.

Instead of trying more complicated regression models we directly estimate the average education of children under correlated matching. Table 3 gives the average level for selected values of ρ (τ is set equal to zero throughout). The figure reports the same levels and also gives the error bands. The standard errors are computed by the delta method (see Appendix A.6). Note that in this application we have no Z variables, i.e. we assume rather unrealistically that the status quo assignment is not selective.

8 Conclusions

[TO BE COMPLETED]

Figure 1: Average years of education child given correlated sorting; 95% error bands

Table 3: Average education given correlated (ρ) sorting

ρ	$\hat{\beta}_{cs}$	Std($\hat{\beta}_{cs}$)
-.99	11.5	.069
-.8	11.7	.048
-.6	11.9	.040
-.4	12.1	.037
-.2	12.4	.034
0.	12.6	.033
.2	12.8	.031
.4	12.9	.030
.6	13.0	.029
.8	13.0	.029
.99	13.1	.039

References

ANGRIST, J. D. AND A. B. KRUEGER. (1999). "Empirical Strategies in Labor Economics," *Handbook of Labor Economics 3A*: 1277 - xxxx (O. Ashenfelter and D. Card, Eds). New York: Elsevier Science.

ATHEY, S., AND G. IMBENS. (2005). "Identification and Inference in Nonlinear Difference-In-Differences Models," forthcoming *Econometrica*.

ATHEY, S., AND S. STERN. (1998). "An Empirical Framework for Testing Theories About Complementarity in Organizational Design", *NBER Working Paper No. 6600.*

BECKER, GARY S. AND KEVIN M. MURPHY. (2000). *Social Economics: Market Behavior in a Social Environment.* Cambridge, MA: Harvard University Press.

BROCK, W. AND S. DURLAUF. (2001). "Interactions-based Models," *Handbook of Econometrics 5*: 3297 - 3380 (J. Heckman & E. Leamer, Eds.). Amsterdam: North-Holland.

CARD, D., AND A. KRUEGER. (1992). "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States, "*Journal of Political Economy* 100 (1): 1 - 40.

DEHEJIA, R. (2005). "Program evaluation as a decision problem," *Journal of Econometrics* 125 (1-2): 141 - 173.

FLORES, C. (2005). "Estimation of Dose-Response Functions and Optimal Doses with a Continuous Treatment," *Mimeo.*

GRAHAM, B., G. IMBENS, AND G. RIDDER. (2006a). "Complementarity and the Optimal Allocation of Inputs," *Mimeo.*

GRAHAM, B., G. IMBENS, AND G. RIDDER. (2006b). "Measuring the Average Outcome and Inequality Effects of Segregation in the Presence of Social Spillovers," *Mimeo.*

Hardle, W. and T.M. Stoker. (1989). "Investigating smooth multiple regression by the method of average derivatives," *Journal of the American Statistical Association* 84 (408): 986 - 995.

Heckman, J., R. Lalonde, and J. Smith. (2000). "The Economics and Econometrics of Active Labor Markets Programs," *Handbook of Labor Economics 3A*: xxxx - xxxx (O. Ashenfelter and D. Card, Eds). New York: Elsevier Science.

Heckman, J., J. Smith, and N. Clements. (1997). "Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts," *Review of Economic Studies* 64 (4): 487 - 535.

Hirano, K. and J. Porter. (2005). "Asymptotics for statistical treatment rules," *Mimeo.*

Imbens, G. (2000). "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika* 87 (3): 706 - 710.

Imbens, G. (2004). "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Survey," *Review of Economics and Statistics* 86 (1): 4 - 30.

Kremer, m. (1997). "How much does sorting increase inequality," *Quarterly Journal of Economics* 112 (1): 115 - 139.

Linton, O., and J. Nielsen. (1995), "A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration," *Biometrika* 82 (1): 93 - 100.

Manski, C. (1990). "Nonparametric Bounds on Treatment Effects," *American Economic Review* 80 (2): 319 - 323.

Manski, C. (1993). "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies* 60 (3): 531 - 542.

Manski, C. (2003). *Partial Identification of Probability Distributions*. New York: Springer-Verlag.

Manski, C. (2004). "Statistical Treatment Rules for Heterogenous Populations," *Econometrica* 72 (4): 1221 - 1246.

Newey, W. (1994). "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory* 10 (2): 233 - 253.

Powell, J., J. Stock and T. Stoker. (1989). "Semiparametric Estimation of Index Coefficients," *Econometrica* 57 (6): 1403 - 1430.

Rosenbaum, P., and D. Rubin. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70 (1): 41 - 55.

Legros, P. and A. Newman. (2004). "Beauty is a beast, frog is a prince: assortative matching with nontransferabilities," Mimeo.

Sacerdote, B. (2001). "Peer effects with random assignment: results for Dartmouth roommates," *Quarterly Journal of Economics* 116 (2): 681 - 704.

Notation

$b_N = N^\alpha$ is bandwidth

$K = \dim(Z)$

$W_\lambda = \lambda \cdot X + (1-\lambda) \cdot W$ for testing ARE's

β's are average outcomes under various policies

β^{pam} for positive assortative matching

β^{nam} for negative assortative matching

β^{rm} for random matching

β^{sq} for status quo

$\beta^{\text{cm}}(\rho, \tau)$ for correlated matching

γ is limit of the local complementarity test statistic

$g(w,x,z) = \mathbb{E}[Y|W=w, X=x, Z=z] = h_2(w,x,z)/h_1(w,x,z)$

$m(w,z) = \mathbb{E}[X|W=w, Z=z]$

$K_b(u) = \frac{1}{b_N^{k+2}}\mathcal{K}(u/\sigma)$ is kernel, where the dimension of u is $k+2$. Kernel is bounded, with bounded support $\mathbb{U} \subset \mathbb{R}^{k+2}$, and of order s.

Support of random variable Z is $\mathcal{Z}$

$V = (W, X, Z')'$ is collection of all random right hand side variables

N observations, (Y_i, V_i) $i = 1, \ldots, N$.

Appendices

A Proofs of Theorems

A.1 Proof of Theorem 4.1

$$d(w_j, x_k) = \frac{\phi_c(\Phi_c^{-1}(F_W(w_j)), \Phi_c^{-1}(F_X(x_k)); \rho)}{\phi_c(\Phi_c^{-1}(F_W(w_j)))\phi_c(\Phi_c^{-1}(F_X(x_k)))}$$

$$\hat{d}(w_j, x_k) = \frac{\phi_c(\Phi_c^{-1}(\hat{F}_W(w_j)), \Phi_c^{-1}(\hat{F}_X(x_k)); \rho)}{\phi_c(\Phi_c^{-1}(\hat{F}_W(w_j)))\phi_c(\Phi_c^{-1}(\hat{F}_X(x_k)))}$$

$$h_1(w, x, z) = \int y f_{YWXZ}(y, w, x, z)\mathrm{d}y \tag{1.26}$$

$$h_2(w, x, z) = f_{WXZ}(w, x, z) \tag{1.27}$$

The ARE for correlated sorting is

$$\hat{\beta}_{CS} = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\hat{g}(w_j, x_k, z_k)\hat{d}(w_j, x_k)$$

with

$$\hat{g}(w, x, z) = \frac{\hat{h}_1(w, x, z)}{\hat{h}_2(w, x, z)}$$

$$g(w, x, z) = \frac{h_1(w, x, z)}{h_2(w, x, z)}$$

Using the identity

$$\hat{a}\hat{b} = ab + b(\hat{a} - a) + a(\hat{b} - b) + (\hat{a} - a)(\hat{b} - b) \tag{1.28}$$

we have

$$\hat{\beta}_{CS} = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}g(w_j, x_k, z_k)d(w_j, x_k)+ \tag{1.29}$$

$$\frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}(\hat{g}(w_j, x_k, z_k) - g(w_j, x_k, z_k))d(w_j, x_k)+ \tag{1.30}$$

$$\frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}g(w_j, x_k, z_k)(\hat{d}(w_j, x_k) - d(w_j, x_k))+ \tag{1.31}$$

$$\frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}(\hat{g}(w_j, x_k, z_k) - g(w_j, x_k, z_k))(\hat{d}(w_j, x_k) - d(w_j, x_k)) \tag{1.32}$$

In this decomposition the main term (1.29) is a two-sample U-statistic, the terms (1.30) and (1.31) are correction terms that account for the estimation of the regression function and the CDFs, and (1.32) is a remainder term that will be shown to be $o_p(1)$.

To obtain an asymptotically linear representation we express (1.30) and (1.31) as averages. We first consider (1.30). Using

$$\frac{\hat{a}}{\hat{b}} = \frac{a}{b} + \frac{1}{b}(\hat{a}-a) - \frac{a}{b^2}(\hat{b}-b) + \frac{a}{b^2\hat{b}}(\hat{b}-b)^2 - \frac{1}{b\hat{b}}(\hat{a}-a)(\hat{b}-b) \tag{1.33}$$

we obtain

$$\hat{g}(w,x,z) - g(w,x,z) = \frac{1}{h_2(w,x,z)}(\hat{h}_1(w,x,z) - g(w,x,z)\hat{h}_2(w,x,z)) +$$

$$\frac{h_1(w,x,z)}{h_2(w,x,z)\hat{h}_2(w,x,z)}(\hat{h}_2(w,x,z) - h_2(w,x,z))^2 -$$

$$\frac{1}{h_2(w,x,z)\hat{h}_2(w,x,z)}(\hat{h}_1(w,x,z) - h_1(w,x,z))(\hat{h}_2(w,x,z) - h_2(w,x,z))$$

Substitution in (1.30) gives

$$\frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\frac{1}{h_2(w_j,x_k,z_k)}(\hat{h}_1(w_j,x_k,z_k) - g(w_j,x_k,z_k)\hat{h}_2(w_j,x_k,z_k))d(w_j,x_k)+ \tag{1.34}$$

$$\frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\frac{h_1(w_j,x_k,z_k)d(w_j,x_k)}{h_2(w_j,x_k,z_k)\hat{h}_2(w_j,x_k,z_k)}(\hat{h}_2(w_j,x_k,z_k) - h_2(w_j,x_k,z_k))^2+ \tag{1.35}$$

$$\frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\frac{d(w_j,x_k)}{h_2(w_j,x_k,z_k)\hat{h}_2(w_j,x_k,z_k)}(\hat{h}_1(w_j,x_k,z_k) - h_1(w_j,x_k,z_k))(\hat{h}_2(w_j,x_k,z_k) - h_2(w_j,x_k,z_k)) \tag{1.36}$$

The main term (1.34) will be expressed as an average. The terms (1.35) and (1.36) are remainder terms that will be shown to be $o_p(1)$.

Substitution of $\hat{h}_1$ and $\hat{h}_2$ gives the following expression for (1.34)

$$\frac{1}{N^3}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\frac{d(w_j,x_k)}{f_{WXZ}(w_j,x_k,z_k)}(y_i - g(w_j,x_k,z_k))K_\sigma(w_j-w_i,x_k-x_i,z_k-z_i) \tag{1.37}$$

This is a V-statistic with kernel

$$h(v_i,v_j,v_k) = \frac{d(w_j,x_k)}{f_{WXZ}(w_j,x_k,z_k)}(y_i - g(w_j,x_k,z_k))K_\sigma(w_j-w_i,x_k-x_i,z_k-z_i)$$

with $v = (w\ x\ z')'$. To apply the V-statistic projection theorem to (1.37) we need that

$$\mathrm{E}(h(v_i,v_j,v_k)^2) < \infty$$

Because $d(w_j, x_k)$ is obviously bounded and by assumption 2 $f_{WXZ}(w_j, x_k, z_k)$ is bounded from 0, this holds if

$$\mathrm{E}\left[(y_i - g(w_j, x_k, z_k))^2 K_\sigma(w_j - w_i, x_k - x_i, z_k - z_i)^2\right] < \infty \tag{1.38}$$

This is true because

$$\mathrm{E}\left[y_i^2 K_\sigma(w_j - w_i, x_k - x_i, z_k - z_i)^2\right] =$$

$$\mathrm{E}_{w_j,x_k,z_k}\left(\int \mathrm{E}(y^2|w, x, z) K_\sigma(w_j - w, x_k - x, z_k - z)^2 f_{WXZ}(w, x, z)\mathrm{d}w\mathrm{d}x\mathrm{d}z\right) =$$

$$\mathrm{E}_{w_j,x_k,z_k}\left(\int \mathrm{E}(y^2|w_j - r\sigma, x_k - s\sigma, z_k - t\sigma) K(r, s, t)^2 f_{WXZ}(w_j - r\sigma, x_k - s\sigma, z_k - t\sigma)\mathrm{d}r\mathrm{d}s\mathrm{d}t\right) \leq$$

$$C\mathrm{E}(y^2) < \infty$$

because $\mathrm{E}(y^2)$ and K are bounded by assumptions 1 and 4. Here and in the sequel we use a change of variables in the integral to $r = (w_j - w)/\sigma, s = (x_k - x)/\sigma, t = (z - z_k)/\sigma$. For (1.38) we also need that (using the same change of variables in the integral)

$$\left|\mathrm{E}\left[y_i g(w_j, x_k, z_k) K_\sigma(w_j - w_i, x_k - x_i, z)k - z_i)^2\right]\right| \leq$$

$$\mathrm{E}_{w_j,x_k,z_k}\left[|g(w_j, x_k, z_k)| \int |g(w, x, z)| K_\sigma(w_j - w, x_k - x, z_k - z)^2 f_{WXZ}(w, x, z)\mathrm{d}w\mathrm{d}x\mathrm{d}z\right] \leq$$

$$\mathrm{E}_{w_j,x_k,z_k}(|g(w_j, x_k, z_k)|)\mathrm{E}_{w_i,x_i,z_i}(|g(w_i, x_i, z_i)|) < \infty$$

by assumptions 1-3. Finally, for (1.38) we need

$$\mathrm{E}\left[g(w_j, x_k, z_k)^2 K_\sigma(w_j - w_i, x_k - x_i, z_k - z_i)^2\right] \leq C\mathrm{E}_{w_j,x_k,z_k}(g(w_j, x_k, z_k)^2) < \infty$$

by assumptions 1-3.

Applying the V-statistic projection to (1.37) we have that

$$\frac{1}{N^3}\sum_{i=1}^N\sum_{j=1}^N\sum_{k=1}^N \frac{d(w_j, x_k)}{f_{WXZ}(w_j, x_k, z_k)}(y_i - g(w_j, x_k, z_k))K_\sigma(w_j - w_i, x_k - x_i, z_k - z_i) =$$

$$\frac{1}{N}\sum_{i=1}^N \int \frac{d(w_j, x_k)}{f_{WXZ}(w_j, x_k, z_k)}(y_i - g(w_j, x_k, z_k))K_\sigma(w_j - w_i, x_k - x_i, z_k - z_i) f_W(w_j) f_{XZ}(x_k, z_k)\mathrm{d}w_j\mathrm{d}x_k\mathrm{d}z_k + \tag{1.39}$$

$$\frac{1}{N}\sum_{j=1}^N \int \frac{d(w_j, x_k)}{f_{WXZ}(w_j, x_k, z_k)}(g(w_i, x_i, z_i) - g(w_j, x_k, z_k))K_\sigma(w_j - w_i, x_k - x_i, z_k - z_i)\cdot \tag{1.40}$$

$$f_{XZ}(x_k, z_k) f_{WXZ}(w_i, x_i, z_i)\mathrm{d}x_k\mathrm{d}z_k\mathrm{d}w_i\mathrm{d}x_i\mathrm{d}z_i +$$

$$\frac{1}{N}\sum_{k=1}^N \int \frac{d(w_j, x_k)}{f_{WXZ}(w_j, x_k, z_k)}(g(w_i, x_i, z_i) - g(w_j, x_k, z_k))K_\sigma(w_j - w_i, x_k - x_i, z_k - z_i)\cdot \tag{1.41}$$

$$f_W(w_j)f_{WXZ}(w_i,x_i,z_i)\mathrm{d}w_j\mathrm{d}w_i\mathrm{d}x_i\mathrm{d}z_i + o_p\left(\frac{1}{\sqrt{N}}\right)$$

We show that only the first term (1.39) has a contribution to the asymptotic distribution. Define for some sequence $\sigma_N \downarrow 0$

$$a_{iN} = \int \frac{d(w_j,x_k)}{f_{WXZ}(w_j,x_k,z_k)}(y_i - g(w_j,x_k,z_k))K_{\sigma_N}(w_j - w_i, x_k - x_i, z_k - z_i)f_W(w_j)f_{XZ}(x_k,z_k)\mathrm{d}w_j\mathrm{d}x_k\mathrm{d}z_k$$

Then

$$|\sqrt{N}\mathrm{E}(a_{iN})| = \left|\sqrt{N}\mathrm{E}_{w_i,x_i,z_i}\left[\int \frac{d(w_j,x_k)}{f_{WXZ}(w_j,x_k,z_k)}(g(w_i,x_i,z_i) - g(w_j,x_k,z_k))\cdot\right.\right.$$

$$K_{\sigma_N}(w_j - w_i, x_k - x_i, z_k - z_i)f_W(w_j)f_{XZ}(x_k,z_k)\mathrm{d}w_j\mathrm{d}x_k\mathrm{d}z_k]| =$$

$$\left|\int \frac{d(w_j,x_k)}{f_{WXZ}(w_j,x_k,z_k)}\sqrt{N}(g(w_j - r\sigma_N, x_k - s\sigma_N, z_k - t\sigma_N) - g(w_j,x_k,z_k))K(r,s,t)\cdot\right.$$

$$f_W(w_j)f_{XZ}(x_k,z_k)f_{WXZ}(w_j - r\sigma_N, x_k - s\sigma_N, z_k - t\sigma_N)\mathrm{d}r\mathrm{d}s\mathrm{d}t\mathrm{d}w_j\mathrm{d}x_k\mathrm{d}z_k|$$

By Taylor's theorem

$$g(w_j - r\sigma_N, x_k - s\sigma_N, z_k - t\sigma_N) - g(w_j,x_k,z_k) = \sum_{j=0}^{S-1}\frac{(-\sigma_N)^{j(K+2)}}{j!}\left(r\frac{\partial}{\partial w} + s\frac{\partial}{\partial x} + t'\frac{\partial}{\partial z}\right)^j g(w_j,x_k,z_k) +$$

$$\sigma_N^{S(K+2)}\frac{(-1)^{S(K+2)}}{S!}\left(r\frac{\partial}{\partial w} + s\frac{\partial}{\partial x} + t'\frac{\partial}{\partial z}\right)^S g(w_j - \overline{r}\sigma_N, x_k - \overline{s}\sigma_N, z_k - \overline{t}\sigma_N)$$

with $\overline{r},\overline{s},\overline{t}$ intermediate between 0 and r,s,t. Because g is assumed to be S times differentiable with an S-th derivative that is bounded on $\Re^{K+2}$ and the kernel is of order S, we have that

$$|\sqrt{N}\mathrm{E}(a_{iN})| \le C\sqrt{N}\sigma_N^{S(K+2)}$$

where we also use assumptions 1,2.

We have by assumptions 1-3 and using the same change of variables as before

$$|a_{iN}| \le C\int |y_i - g(w_j,x_k,z_k)||K_{\sigma_N}(w_j - w_i, x_k - x_i, z_k - z_i)|f_W(w_j)f_{XZ}(x_k,z_k)\mathrm{d}w_j\mathrm{d}x_k\mathrm{d}z_k \le$$

$$C|y_i|\int |K_{\sigma_N}(w_j - w_i, x_k - x_i, z_k - z_i)|f_W(w_j)f_{XZ}(x_k,z_k)\mathrm{d}w_j\mathrm{d}x_k\mathrm{d}z_k +$$

$$C\int |g(w_j,x_k,z_k)||K_{\sigma_N}(w_j - w_i, x_k - x_i, z_k - z_i)|f_W(w_j)f_{XZ}(x_k,z_k)\mathrm{d}w_j\mathrm{d}x_k\mathrm{d}z_k \le$$

$$C_1|y_i| + C_2\mathrm{E}_{w_j,x_k,z_k}(|g(w_j,x_k,z_k)|)$$

and the right-hand side has a finite expected value. The upper bound on $|a_{iN}|$ also yields an upper bound on a_{iN}^2 that has a finite expected value. We have

$$\mathrm{E}(a_{iN}^2) = \mathrm{E}_{y_i,w_i,x_i,z_i}\left[\int\int \frac{d(w_j,x_k)}{f_{WXZ}(w_j,x_k,z_k)}\frac{d(w_j',x_k')}{f_{WXZ}(w_j',x_k',z_k')}(y_i - g(w_j,x_k,z_k))(y_i - g(w_j',x_k',z_k'))\cdot\right.$$

$$K_{\sigma_N}(w_j - w_i, x_k - x_i, z_k - z_i) K_{\sigma_N}(w_j' - w_i, x_k' - x_i, z_k' - z_i) \cdot$$

$$f_W(w_j) f_{XZ}(x_k, z_k) f_W(w_j') f_{XZ}(x_k', z_k') \mathrm{d}w_j \mathrm{d}x_k \mathrm{d}z_k \mathrm{d}w_j' \mathrm{d}x_k' \mathrm{d}z_k' \Big] =$$

$$\mathrm{E}_{y_i,w_i,x_i,z_i} \Bigg[\int\int \frac{d(w_i + r\sigma_N, x_i + s\sigma_N)}{f_{WXZ}(w_i + r\sigma_N, x_i + s\sigma_N, z_i + t\sigma_N)} \frac{d(w_i + r'\sigma_N, x_i + s'\sigma_N)}{f_{WXZ}(w_i + r'\sigma_N, x_i + s'\sigma_N, z_i + t'\sigma_N)} \cdot$$

$$(y_i - g(w_i + r\sigma_N, x_i + s\sigma_N, z_i + t\sigma_N))(y_i - g(w_i + r'\sigma_N, x_i + s'\sigma_N, z_i + t'\sigma_N)) \cdot$$

$$K(r,s,t) K(r',s',t') f_W(w_i + r\sigma_N) f_{XZ}(x_i + s\sigma_N, z_i + t\sigma_N) \cdot$$

$$f_W(w_i + r'\sigma_N) f_{XZ}(x_i + s'\sigma_N, z_i + t'\sigma_N) \mathrm{d}r \mathrm{d}s \mathrm{d}t \mathrm{d}r' \mathrm{d}s' \mathrm{d}t'$$

Hence by dominated convergence

$$\lim_{N\to\infty} \mathrm{E}(a_{iN}^2) = \mathrm{E}_{y_i,w_i,x_i,z_i} \left[\frac{d(w_i,x_i)^2}{f_{WXZ}(w_i,x_i,z_i)^2} (y_i - g(w_i,x_i,z_i))^2 f_W(w_i)^2 f_{XZ}(x_i,z_i)^2 \right] \equiv V$$

We conclude that for any sequence σ_N that satisfies $\sqrt{N}\sigma_N^{S(K+2)} \downarrow 0$

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N (a_{iN} - \mathrm{E}(a_{iN})) = \frac{1}{\sqrt{N}} \sum_{i=1}^N a_{iN} + o_p(1)$$

so that by the Lindeberg-Feller Central Limit Theorem

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N a_{iN} \xrightarrow{d} N(0,V) \tag{1.42}$$

Next we consider (1.40). With a change of variables and by the same assumptions that were used to bound $|\sqrt{N}\mathrm{E}(a_{iN})|$ we have

$$\mathrm{E}\left[\left| \frac{1}{\sqrt{N}} \sum_{j=1}^N \int \frac{d(w_j,x_k)}{f_{WXZ}(w_j,x_k,z_k)} (g(w_i,x_i,z_i) - g(w_j,x_k,z_k)) K_\sigma(w_j - w_i, x_k - x_i, z_k - z_i) \cdot\right.\right.$$

$$f_{XZ}(x_k,z_k) f_{WXZ}(w_i,x_i,z_i) \mathrm{d}x_k \mathrm{d}z_k \mathrm{d}w_i \mathrm{d}x_i \mathrm{d}z_i |] \leq$$

$$\sqrt{N}\mathrm{E}_{w_j}\left[\left| \int \frac{d(w_j,x_k)}{f_{WXZ}(w_j,x_k,z_k)} (g(w_j - r\sigma_N, x_k - s\sigma_N, z_k - t\sigma_N) - g(w_j,x_k,z_k)) K(r,s,t) \cdot\right.\right.$$

$$f_{XZ}(x_k,z_k) f_{WXZ}(w_j - r\sigma_N, x_k - s\sigma_N, z_k - t\sigma_N) \mathrm{d}x_k \mathrm{d}z_k \mathrm{d}r \mathrm{d}s \mathrm{d}t |] \leq$$

$$C\sqrt{N}\sigma_N^{S(K+2)} \int \left| \left(r\frac{\partial}{\partial w} + s\frac{\partial}{\partial x} + t'\frac{\partial}{\partial z} \right)^S g(w_j - \overline{r}\sigma_N, x_k - \overline{s}\sigma_N, z_k - \overline{t}\sigma_N) \right| |K(r,s,t)| \cdot$$

$$f_W(w_j) f_{XZ}(x_k,z_k) \mathrm{d}w_j \mathrm{d}x_k \mathrm{d}z_k \mathrm{d}r \mathrm{d}s \mathrm{d}t$$

so that if $\sqrt{N}\sigma_N^{S(K+2)} \downarrow 0$

$$\frac{1}{\sqrt{N}} \sum_{j=1}^N \int \frac{d(w_j,x_k)}{f_{WXZ}(w_j,x_k,z_k)} (g(w_i,x_i,z_i) - g(w_j,x_k,z_k)) K_\sigma(w_j - w_i, x_k - x_i, z_k - z_i) \cdot \tag{1.43}$$

$$f_{XZ}(x_k,z_k) f_{WXZ}(w_i,x_i,z_i) \mathrm{d}x_k \mathrm{d}z_k \mathrm{d}w_i \mathrm{d}x_i \mathrm{d}z_i = o_p(1)$$

by the Markov inequality. An analogous argument gives that

$$\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\int\frac{d(w_j,x_k)}{f_{WXZ}(w_j,x_k,z_k)}(g(w_i,x_i,z_i)-g(w_j,x_k,z_k))K_\sigma(w_j-w_i,x_k-x_i,z_k-z_i)\cdot \tag{1.44}$$

$$f_W(w_j)f_{WXZ}(w_i,x_i,z_i)\mathrm{d}w_j\mathrm{d}w_i\mathrm{d}x_i\mathrm{d}z_i=o_p(1)$$

if $\sqrt{N}\sigma_N^{S(K+2)}\downarrow 0$.

Next, we consider (1.35). Newey (1994) shows that under assumptions 1-4

$$\sup_{w,x,z\in\mathcal{W}\times\mathcal{X}\times\mathcal{Z}}|\hat{h}_j(w,x,z)-h_j(w,x,z)|=O_p\left(\sqrt{\frac{\ln N}{N}}\sigma_N^{-\frac{K}{2}-1}+\sigma_N^{S(K+2)}\right) \tag{1.45}$$

for $j=1,2$. Hence we have

$$\left|\frac{\sqrt{N}}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\frac{h_1(w_j,x_k,z_k)d(w_j,x_k)}{h_2(w_j,x_k,z_k)\hat{h}_2(w_j,x_k,z_k)}(\hat{h}_2(w_j,x_k,z_k)-h_2(w_j,x_k,z_k))^2\right|\leq$$

$$\frac{\left(N^{\frac{1}{4}}\sup_{w,x,z\in\mathcal{W}\times\mathcal{X}\times\mathcal{Z}}|\hat{h}_2(w,x,z)-h_2(w,x,z)|\right)^2}{\inf_{w,x,z\in\mathcal{W}\times\mathcal{X}\times\mathcal{Z}}|\hat{h}_2(w,x,z)|}\frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\left|\frac{h_1(w_j,x_k,z_k)d(w_j,x_k)}{h_2(w_j,x_k,z_k)}\right|$$

Now

$$\inf_{w,x,z\in\mathcal{W}\times\mathcal{X}\times\mathcal{Z}}|\hat{h}_2(w,x,z)|\geq\inf_{w,x,z\in\mathcal{W}\times\mathcal{X}\times\mathcal{Z}}|h_2(w,x,z)|-\sup_{w,x,z\in\mathcal{W}\times\mathcal{X}\times\mathcal{Z}}|\hat{h}_2(w,x,z)-h_2(w,x,z)|$$

Hence

$$\left|\frac{\sqrt{N}}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\frac{h_1(w_j,x_k,z_k)d(w_j,x_k)}{h_2(w_j,x_k,z_k)\hat{h}_2(w_j,x_k,z_k)}(\hat{h}_2(w_j,x_k,z_k)-h_2(w_j,x_k,z_k))^2\right|=o_p(1) \tag{1.46}$$

if σ_N is such that

$$\frac{N^{\frac{1}{4}}}{\sqrt{\ln N}}\sigma_N^{\frac{K}{2}+1}\to\infty \qquad N^{\frac{1}{4}}\sigma_N^{S(K+2)}\to 0 \tag{1.47}$$

An analogous argument shows that for (1.36) we have

$$\frac{\sqrt{N}}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\frac{d(w_j,x_k)}{h_2(w_j,x_k,z_k)\hat{h}_2(w_j,x_k,z_k)}(\hat{h}_1(w_j,x_k,z_k)-h_1(w_j,x_k,z_k))(\hat{h}_2(w_j,x_k,z_k)-h_2(w_j,x_k,z_k))= \tag{1.48}$$

$$o_p(1)$$

if σ_N satisfies (1.47). This completes the discussion of the first part of the correction term.

The second part of the correction term is obtained from (1.31). Using the definition of k in (1.58) and a second-order Taylor expansion we have

$$\hat{d}(w_j,x_k)-d(w_j,x_k)=k(\hat{F}_W(w_j),\hat{F}_X(x_k))-k(F_W(w_j),F_X(x_k))=$$

$$\frac{\partial k}{\partial s_1}(F_W(w_j), F_X(x_k))(\hat{F}_W(w_j) - F_W(w_j)) + \frac{\partial k}{\partial s_2}(F_W(w_j), F_X(x_k))(\hat{F}_X(x_k) - F_X(x_k)) +$$

$$\frac{1}{2}\frac{\partial^2 k}{\partial s_1^2}(\overline{F}_W(w_j), \overline{F}_X(x_k))(\hat{F}_W(w_j) - F_W(w_j))^2 + \frac{1}{2}\frac{\partial^2 k}{\partial s_2^2}(F_W(w_j), F_X(x_k))(\hat{F}_X(x_k) - F_X(x_k))^2 +$$

$$\frac{\partial^2 k}{\partial s_1 \partial s_2}(\overline{F}_W(w_j), \overline{F}_X(x_k))(\hat{F}_W(w_j) - F_W(w_j))(\hat{F}_X(x_k) - F_X(x_k))$$

with $\overline{F}_W(w_j), \overline{F}_X(x_k)$ intermediate values. Substitution in (1.31) gives

$$\frac{\sqrt{N}}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} g(w_j, x_k, z_k)(\hat{d}(w_j, x_k) - d(w_j, x_k)) =$$

$$\frac{\sqrt{N}}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} g(w_j, x_k, z_k)\frac{\partial k}{\partial s_1}(F_W(w_j), F_X(x_k))(\hat{F}_W(w_j) - F_W(w_j)) + \tag{1.49}$$

$$\frac{\sqrt{N}}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} g(w_j, x_k, z_k)\frac{\partial k}{\partial s_2}(F_W(w_j), F_X(x_k))(\hat{F}_X(x_k) - F_X(x_k)) + \tag{1.50}$$

$$\frac{1}{2}\frac{\sqrt{N}}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} g(w_j, x_k, z_k)\frac{\partial^2 k}{\partial s_1^2}(\overline{F}_W(w_j), \overline{F}_X(x_k))(\hat{F}_W(w_j) - F_W(w_j))^2 + \tag{1.51}$$

$$\frac{1}{2}\frac{\sqrt{N}}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} g(w_j, x_k, z_k)\frac{\partial^2 k}{\partial s_2^2}(F_W(w_j), F_X(x_k))(\hat{F}_X(x_k) - F_X(x_k))^2 + \tag{1.52}$$

$$\frac{\sqrt{N}}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} g(w_j, x_k, z_k)\frac{\partial^2 k}{\partial s_1 \partial s_2}(\overline{F}_W(w_j), \overline{F}_X(x_k))(\hat{F}_W(w_j) - F_W(w_j))(\hat{F}_X(x_k) - F_X(x_k)) \tag{1.53}$$

where the derivatives are given in (1.59)-(1.63).

The leading terms (1.49) and (1.50) are V-statistics. For instance, (1.49) can expressed as

$$\frac{\sqrt{N}}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N} g(w_j, x_k, z_k)\frac{\partial k}{\partial s_1}(F_W(w_j), F_X(x_k))(I(w_l \leq w_j) - F_W(w_j)) = \tag{1.54}$$

$$\frac{1}{\sqrt{N}}\sum_{1=1}^{N} \mathrm{E}_{w_j, x_k, z_k}\left[g(w_j, x_k, z_k)\frac{\partial k}{\partial s_1}(F_W(w_j), F_X(x_k))(I(w_i \leq w_j) - F_W(w_j))\right] + o_p(1)$$

and (1.50) is

$$\frac{1}{\sqrt{N}}\sum_{1=1}^{N} \mathrm{E}_{w_j, x_k, z_k}\left[g(w_j, x_k, z_k)\frac{\partial k}{\partial s_2}(F_W(w_j), F_X(x_k))(I(x_i \leq x_k) - F_X(x_k))\right] + o_p(1) \tag{1.55}$$

both by the V-statistic projection theorem.

The terms in (1.51)-(1.53) are $o_p(1)$. We show this for (1.51) with the proof for the other terms being analogous. First, note that inspection of (1.61) shows that this is a bounded function. Hence

$$\left|\frac{1}{2}\frac{\sqrt{N}}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} g(w_j, x_k, z_k)\frac{\partial^2 k}{\partial s_1^2}(\overline{F}_W(w_j), \overline{F}_X(x_k))(\hat{F}_W(w_j) - F_W(w_j))^2\right| \leq$$

$$C\sqrt{N}\sup_{w\in\mathcal{W}}|\hat{F}_W(w) - F_W(w)|^2 \leq C\left(N^{\frac{1}{4}}\sup_{w}|\hat{F}_W(w) - F_W(w)|\right)^2 = o_p(1)$$

by the law of the iterated logarithm.

The final step in the proof is to deal with (1.32). We have

$$\left|\frac{\sqrt{N}}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}(\hat{g}(w_j, x_k, z_k) - g(w_j, x_k, z_k))(\hat{d}(w_j, x_k) - d(w_j, x_k))\right| \leq$$

$$N^{\frac{1}{4}}\sup_{w,x,z\in\mathcal{W}\times\mathcal{X}\times\mathcal{Z}}|\hat{g}(w,x,z) - g(w,x,z)| \cdot N^{\frac{1}{4}}\sup_{w,x\in\mathcal{W}\times\mathcal{X}}|\hat{d}(w,x) - d(w,x)|$$

We have

$$N^{\frac{1}{4}}\sup_{w,x\in\mathcal{W}\times\mathcal{X}}|\hat{d}(w,x) - d(w,x)| =$$

$$N^{\frac{1}{4}}\sup_{w,x\in\mathcal{W}\times\mathcal{X}}\left|\frac{\partial k}{\partial s_1}(\overline{F}_W(w), \overline{F}_X(x))(\hat{F}_W(w) - F_W(w)) + \frac{\partial k}{\partial s_2}(\overline{F}_W(w), \overline{F}_X(x))(\hat{F}_X(x) - F_X(x))\right| \leq$$

$$C_1 N^{\frac{1}{4}}\sup_{w\in\mathcal{W}}|\hat{F}_W(w) - F_W(w)| + C_2 N^{\frac{1}{4}}\sup_{x\in\mathcal{X}}|\hat{F}_X(x) - F_X(x)| = o_p(1)$$

Using (1.33) we have

$$\sup_{w,x,z\in\mathcal{W}\times\mathcal{X}\times\mathcal{Z}}|\hat{g}(w,x,z) - g(w,x,z)| \leq$$

$$\sup_{w,x,z\in\mathcal{W}\times\mathcal{X}\times\mathcal{Z}}\frac{1}{|h_2(w,x,z)|}|\hat{h}_1(w,x,z) - h_1(w,x,z)| + \sup_{w,x,z\in\mathcal{W}\times\mathcal{X}\times\mathcal{Z}}\frac{|g(w,x,z)|}{|h_2(w,x,z)|}|\hat{h}_2(w,x,z) - h_2(w,x,z)| +$$

$$\sup_{w,x,z\in\mathcal{W}\times\mathcal{X}\times\mathcal{Z}}\frac{|g(w,x,z)|}{|h_2(w,x,z)\hat{h}_2(w,x,z)|}|\hat{h}_2(w,x,z) - h_2(w,x,z)|^2 +$$

$$\sup_{w,x,z\in\mathcal{W}\times\mathcal{X}\times\mathcal{Z}}\frac{1}{|h_2(w,x,z)\hat{h}_2(w,x,z)|}|\hat{h}_2(w,x,z) - h_2(w,x,z)||\hat{h}_1(w,x,z) - h_1(w,x,z)|$$

Hence by (1.45) we have that for a sequence σ_N that satisfies (1.47)

$$N^{\frac{1}{4}}\sup_{w,x,z\in\mathcal{W}\times\mathcal{X}\times\mathcal{Z}}|\hat{g}(w,x,z) - g(w,x,z)| = o_p(1)$$

This completes the proof of the theorem. □

A.2 Derivatives of the Truncated Normal Cupola

We consider the functions

$$c_1(s) = \phi_c(\Phi_c^{-1}(s))$$

and

$$c_2(s_1, s_2) = \phi_c(\Phi_c^{-1}(s_1), \Phi_c^{-1}(s_2); \rho)$$

with $0 \leq s \leq 1$ and $0 \leq s_1, s_2 \leq 1$. The derivatives are

$$\frac{\mathrm{d}c_1}{\mathrm{d}s}(s) = -\Phi_c^{-1}(s) \tag{1.56}$$

$$\frac{\partial c_2}{\partial s_1}(s_1, s_2) = -\frac{\phi_c(\Phi_c^{-1}(s_1), \Phi_c^{-1}(s_2); \rho)}{1-\rho^2} \frac{\Phi_c^{-1}(s_1) - \rho\Phi_c^{-1}(s_2)}{\phi_c(\Phi_c^{-1}(s_1))} \tag{1.57}$$

Define

$$k(s_1, s_2) = \frac{c_2(s_1, s_2)}{c_1(s_1)c_2(s_2)} \tag{1.58}$$

We have

$$\frac{\partial k}{\partial s_1}(s_1, s_2) = -\frac{\phi_c(\Phi_c^{-1}(s_1), \Phi_c^{-1}(s_2); \rho)}{\phi_c(\Phi_c^{-1}(s_1))\phi_c(\Phi_c^{-1}(s_2))} \left\{ \frac{\rho^2\Phi_c^{-1}(s_1) + \rho\Phi_c^{-1}(s_2)}{\phi_c(\Phi_c^{-1}(s_1))} \right\} \tag{1.59}$$

$$\frac{\partial k}{\partial s_2}(s_1, s_2) = -\frac{\phi_c(\Phi_c^{-1}(s_2), \Phi_c^{-1}(s_1); \rho)}{\phi_c(\Phi_c^{-1}(s_1))\phi_c(\Phi_c^{-1}(s_2))} \left\{ \frac{\rho^2\Phi_c^{-1}(s_2) + \rho\Phi_c^{-1}(s_1)}{\phi_c(\Phi_c^{-1}(s_2))} \right\} \tag{1.60}$$

$$\frac{\partial^2 k}{\partial s_1^2}(s_1, s_2) = \frac{\phi_c(\Phi_c^{-1}(s_2), \Phi_c^{-1}(s_1); \rho)}{\phi_c(\Phi_c^{-1}(s_1))^3\phi_c(\Phi_c^{-1}(s_2))} \cdot \tag{1.61}$$

$$\left\{ \left(\rho^2\Phi_c^{-1}(s_2) + \rho\Phi_c^{-1}(s_1)\right)^2 - \left(\rho^2\Phi_c^{-1}(s_2) + \rho\Phi_c^{-1}(s_1)\right) \Phi_c^{-1}(s_1) - \rho^2 \right\}$$

$$\frac{\partial^2 k}{\partial s_2^2}(s_1, s_2) = \frac{\phi_c(\Phi_c^{-1}(s_2), \Phi_c^{-1}(s_1); \rho)}{\phi_c(\Phi_c^{-1}(s_1))\phi_c(\Phi_c^{-1}(s_2))^3} \cdot \tag{1.62}$$

$$\left\{ \left(\rho^2\Phi_c^{-1}(s_1) + \rho\Phi_c^{-1}(s_2)\right)^2 - \left(\rho^2\Phi_c^{-1}(s_1) + \rho\Phi_c^{-1}(s_2)\right) \Phi_c^{-1}(s_2) - \rho^2 \right\}$$

$$\frac{\partial^2 k}{\partial s_1 \partial s_2}(s_1, s_2) = \frac{\phi_c(\Phi_c^{-1}(s_2), \Phi_c^{-1}(s_1); \rho)}{\phi_c(\Phi_c^{-1}(s_1))^2\phi_c(\Phi_c^{-1}(s_2))^2} \cdot \tag{1.63}$$

$$\left\{ \left(\rho^2\Phi_c^{-1}(s_1) + \rho\Phi_c^{-1}(s_2)\right) \left(\rho^2\Phi_c^{-1}(s_2) + \rho\Phi_c^{-1}(s_1)\right) - \rho \right\}$$

A.3 Materials related to derivation of large sample properties of β^{pam}

Before proving the main results we present some preliminary results.
For a real valued random variable Y with support $\mathbb{Y} = [\underline{y}, \overline{y}]$, define

$$\hat{F}_Y(y) = \sum_{i=1}^{N} 1\{Y_i \leq y\},$$

and, for $q \in (0, 1]$

$$\hat{F}_Y^{-1}(q) = \min\{y \in \mathbb{Y} : \hat{F}_Y(y) \geq q\},$$

and $\hat{F}_Y^{-1}(0) = \underline{y}$.

First some Lemmas from AI.

Lemma A.1 *For any $\delta < 1/2$, with $\mathbb{Y} = [\underline{y}, \overline{y}]$, and $F_Y(y)$ continuously differentiable on $\mathbb{Y}$,*

$$\sup_{y\in\mathbb{Y}} N^{\delta} \cdot |\hat{F}_Y(y) - F_Y(y)| \xrightarrow{p} 0.$$

Proof: See Lemma A.2 in AI.

Lemma A.2 *For any $\delta < 1/2$, with $\mathbb{Y} = [\underline{y}, \overline{y}]$, and $F_Y(y)$ continuously differentiable on $\mathbb{Y}$,*

$$\sup_{q\in[0,1]} N^{\delta} \cdot |\hat{F}_Y^{-1}(q) - F_Y^{-1}(q)| \xrightarrow{p} 0.$$

Proof: See Lemma A.3 in AI.

We will use the following generalization of this:

Lemma A.3 *For any $\delta < 1/2$, with $\mathbb{Y} = [\underline{y}, \overline{y}]$, $\mathbb{X} = [\underline{x}, \overline{x}]$ and $F_Y(y)$ and $F_X(x)$ continuously differentiable on $\mathbb{Y}$, with $F_Y(y)$ bounded away from zero on $\mathbb{Y}$:*

$$\sup_{x\in\mathbb{X}} N^{\delta} \cdot \left|\hat{F}_Y^{-1}\left(\hat{F}_X(x)\right) - F_Y^{-1}\left(F_X(x)\right)\right| \xrightarrow{p} 0.$$

Proof: By the triangle inequality

$$\sup_{x\in\mathbb{X}} N^{\delta} \cdot \left|\hat{F}_Y^{-1}\left(\hat{F}_X(x)\right) - F_Y^{-1}\left(F_X(x)\right)\right|$$
$$\leq \sup_{x\in\mathbb{X}} N^{\delta} \cdot \left|\hat{F}_Y^{-1}\left(\hat{F}_X(x)\right) - F_Y^{-1}\left(\hat{F}_X(x)\right)\right|$$
$$+ \sup_{x\in\mathbb{X}} N^{\delta} \cdot \left|F_Y^{-1}\left(\hat{F}_X(x)\right) - F_Y^{-1}\left(\hat{F}_X(x)\right)\right|$$
$$\leq \sup_{q\in[0,1]} N^{\delta} \cdot \left|\hat{F}_Y^{-1}(q) - F_Y^{-1}(q)\right|$$
$$+ \sup_{x\in\mathbb{X}, y\in\mathbb{Y}} N^{\delta} \cdot \frac{1}{f_Y(y)} \left|\hat{F}_X(x) - F_X(x)\right|.$$

The first term is $o_p(1)$ by Lemma A.2, and the second by the fact that $F_Y(y)$ is continuous differentiable with its derivative bounded away from zero, in combination with Lemma A.1. □

Lemma A.4 (Uniform Convergence) *Suppose $\mathbb{Y} = [\underline{y}, \overline{y}]$, and $F_Y(y)$ is twice continuously differentiable on $\mathbb{Y}$, with its first derivative $f_Y(y) = \frac{\partial F_Y}{\partial y}(y)$ bounded away from zero on $\mathbb{Y}$. Then, for $0 < \eta < 3/4$ and $\delta > \max(2\eta - 1, \eta/2)$,*

$$\sup_{y\in\mathbb{Y}, x\leq N^{-\delta}, x+y\in\mathbb{Y}} N^{\eta}\cdot\left|\hat{F}_Y(y+x) - \hat{F}_Y(y) - (F_Y(y+x) - F_Y(y))\right| \xrightarrow{p} 0.$$

Proof: See Lemma A.5 in AI.

Lemma A.5 () *Suppose $\mathbb{Y} = [\underline{y}, \overline{y}]$, and $F_Y(y)$ is twice continuously differentiable on $\mathbb{Y}$, with its first derivative $f_Y(y) = \frac{\partial F_Y}{\partial y}(y)$ bounded away from zero on $\mathbb{Y}$. Then, for $0 < \eta < 3/4$ and $\delta > \max(2\eta - 1, \eta/2)$,*

$$\sup_{y\in\mathbb{Y}, x\leq N^{-\delta}, x+y\in\mathbb{Y}} N^{\eta}\cdot\left|\hat{F}_Y(y+x) - \hat{F}_Y(y) - f_Y(y)\cdot x\right| \xrightarrow{p} 0.$$

Proof: By the triangle inequality

$$\sup_{y\in\mathbb{Y}, x\leq N^{-\delta}, x+y\in\mathbb{Y}} N^{\eta}\cdot\left|\hat{F}_Y(y+x) - \hat{F}_Y(y) - f_Y(y)\cdot x\right|$$

$$\leq \sup_{y\in\mathbb{Y}, x\leq N^{-\delta}, x+y\in\mathbb{Y}} N^{\eta}\cdot\left|\hat{F}_Y(y+x) - (F_Y(y+x) - F_Y(y))\right|$$
$$+ \sup_{y\in\mathbb{Y}, x\leq N^{-\delta}, x+y\in\mathbb{Y}} N^{\eta}\cdot\left|F_Y(y+x) - F_Y(y) - f_Y(y)\cdot x\right|.$$

The first term on the right-hand side converges to zero in probability by Lemma A.4. To show that the second term converges to zero note that

$$\sup_{y\in\mathbb{Y}, x\leq N^{-\delta}, x+y\in\mathbb{Y}} N^{\eta}\cdot\left|F_Y(y+x) - F_Y(y) - f_Y(y)\cdot x\right|$$

$$\leq \sup_{y\in\mathbb{Y}, x\leq N^{-\delta}, x+y\in\mathbb{Y}, \lambda\in[0,1]} N^{\eta}\cdot\left|f_Y(y+\lambda x)\cdot x - f_Y(y)\cdot x\right|$$

$$\leq \sup_{y\in\mathbb{Y}, x\leq N^{-\delta}, x+y\in\mathbb{Y}} N^{\eta}\cdot\left|F_Y(y+x) - F_Y(y) - f_Y(y)\cdot x\right|$$

$$\leq \sup_{y\in\mathbb{Y}, z\in\mathbb{Y}, x\leq N^{-\delta}, x+y\in\mathbb{Y}} N^{\eta}\cdot\left|\frac{\partial f_Y}{\partial y}(z)\cdot\lambda x^2\right|$$

$$\leq \sup_{y\in\mathbb{Y}, x\leq N^{-\delta}} N^{\eta} x^2 \frac{\partial f_Y}{\partial y}(y) \to 0,$$

because $\frac{\partial f_Y}{\partial y}(y)$ is bounded, $x < N^{-\delta}$, and $\delta > \eta/2$.

Lemma A.6 *Suppose $\mathbb{Y} = [\underline{y}, \overline{y}]$, and $F_Y(y)$ is twice continuously differentiable on $\mathbb{Y}$, with its first derivative $f_Y(y) = \frac{\partial F_Y}{\partial y}(y)$ bounded away from zero on $\mathbb{Y}$. Then, for all $0 < \eta < 5/7$,*

$$\sup_{q\in[0,1]} N^{\eta}\cdot\left|\hat{F}_Y^{-1}(q) - F_Y^{-1}(q) + \frac{1}{f_Y(F_Y^{-1}(q))}\left(\hat{F}_Y(F_Y^{-1}(q)) - q\right)\right| \xrightarrow{p} 0.$$

Proof: See Lemma A.6 in AI.

Next, some results from Newey (1994b).

Lemma A.7 *Suppose that $\mathbb{E}[|Y|^p] < \infty$ for $p > 2$, $\mathbb{E}[|Y|^p|X = x]f_(x)$ is bounded, $\mathbb{X} \subset \mathbb{R}^k$ is compact, Assumption 4.6 is satisfied for $\Delta \geq j$, and $\sigma = \sigma(N)$ such that $\sigma(N)$ is bounded and $N^{1-(2/p)}\sigma(N)^k/\ln(N) \to \infty$. Then*

$$\|\hat{g} - \mathbb{E}[\hat{g}]\|_j = O_p\left(\ln(N)^{1/2}(N\sigma(N)^{k+2j})^{-1/2}\right). \tag{1.64}$$

Lemma A.8 *If Assumption 4.6, 4.7, and 4.8 are satisfied for $d \geq j + s$, then*

$$\|\mathbb{E}[\hat{g}] - g\|_j = O(\sigma(N)^s). \tag{1.65}$$

Lemma A.9 *If the hypotheses of Lemma's A.7 and A.8 are satisfied and Assumption 4.7 is satisfied with for $d \geq j + s$, then*

$$\|\hat{g} - g\|_j = O_p(\ln(N)^{1/2}(N\sigma(N)^{k+2j})^{-1/2} + \sigma(N)^s. \tag{1.66}$$

Next, a result from Newey (1997).

Now we present some results directly relevant to the estimator for β^{pam}.

Lemma A.10

$$\hat{\mu}_{WX} - \hat{\mu}_W = o_p\left(N^{-1/2}\right).$$

Proof: $\hat{\mu}_{WX}$ is a V-statistic. Hence it can be approximated by two sums

$$\hat{\mu}_{WX} = \mathbb{E}[\hat{\mu}_{WX}]$$
$$+\frac{1}{N}\sum_{i=1}^{N}(q_W(W_i) - \mathbb{E}[q_W(W)]) + \frac{1}{N}\sum_{i=1}^{N}(q_X(X_i) - \mathbb{E}[q_X(X)]) + o_p\left(N^{-1/2}\right).$$

However, $q_X(x) = 0$, so that the third term vanishes and

$$\hat{\mu}_{WX} = \mathbb{E}[\hat{\mu}_{WX}] + \frac{1}{N}\sum_{i=1}^{N}(q_W(W_i) - \mathbb{E}[q_W(W)]) + o_p\left(N^{-1/2}\right)$$
$$= \frac{1}{N}\sum_{i=1}^{N}q_W(W_i) + o_p\left(N^{-1/2}\right)$$
$$= \hat{\mu}_W + o_p\left(N^{-1/2}\right)$$

□

Lemma A.11

$$\hat{\mu}_{X_1X_2} - \hat{\mu}_X = o_p\left(N^{-1/2}\right).$$

Proof: The proof is the same as the proof of Lemma A.10, and therefore omitted. □

Lemma A.12

$$\sqrt{N} \cdot \hat{\mu}_W \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}[q_W(W_i)^2]\right),$$

and

$$\sqrt{N} \cdot \hat{\mu}_X \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}[q_X(X_i)^2]\right).$$

Proof: Because $\mathbb{E}[q_{WX}(W,X)] = 0$, it follows that $\mathbb{E}[q_W(W)] = 0$. Then a central limit theorem implies the result. For the second part, we have similarly $\mathbb{E}[q_{X_1X_2}(X, x_2)] = 0$, and therefore $\mathbb{E}[q_X(X)] = 0$, and then a central limit theorem implies the second result. □

Proof of Lemma 4.1:

$$\hat{\theta} - \theta = \frac{1}{N}\sum_{i=1}^{N} \hat{g}\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right), X_i\right) - \mathbb{E}\left[g\left(F_W^{-1}\left(F_X(X)\right), X\right)\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N} \hat{g}\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right), X_i\right) - \frac{1}{N}\sum_{i=1}^{N} g\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right), X_i\right) \tag{1.67}$$

$$- \left(\frac{1}{N}\sum_{i=1}^{N} \hat{g}\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) - \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right)\right) \tag{1.68}$$

$$+ \frac{1}{N}\sum_{i=1}^{N} \hat{g}\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) - \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) \tag{1.69}$$

$$+ \frac{1}{N}\sum_{i=1}^{N} g\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right), X_i\right) - \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(\hat{F}_X(X_i)\right), X_i\right) \tag{1.70}$$

$$- \left(\frac{1}{N}\sum_{i=1}^{N} g\left(\hat{F}_W^{-1}\left(F_X(X_i)\right), X_i\right) - \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right)\right) \tag{1.71}$$

$$+ \frac{1}{N}\sum_{i=1}^{N} g\left(\hat{F}_W^{-1}\left(F_X(X_i)\right), X_i\right) - \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) \tag{1.72}$$

$$+ \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(\hat{F}_X(X_i)\right), X_i\right) - \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) \tag{1.73}$$

$$+ \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) - \mathbb{E}\left[g\left(F_W^{-1}\left(F_X(X)\right), X\right)\right]. \tag{1.74}$$

Since the right-hand side of (4.16) is equal to (1.69), (4.17) equals (1.72), (4.18) equals (1.73), and the first two terms in (4.19) equal (1.74), we only need to show that the sum of (1.67), (1.68), (1.70), and (1.71) are $o_p(N^{-1/2})$. We do this by showing that the sum of (1.67) and (1.68) is $o_p(N^{-1/2})$, and that the sum of (1.70) and (1.71) is $o_p(N^{-1/2})$.

First consider the sum of (1.67) and (1.68).

$$\frac{1}{N}\sum_{i=1}^{N} \hat{g}\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right), X_i\right) - \frac{1}{N}\sum_{i=1}^{N} g\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right), X_i\right)$$

$$-\left(\frac{1}{N}\sum_{i=1}^{N}\hat{g}\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)-\frac{1}{N}\sum_{i=1}^{N}g\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)\right)$$

$$=\frac{1}{N}\sum_{i=1}^{N}\hat{g}\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right),X_i\right)-\frac{1}{N}\sum_{i=1}^{N}\hat{g}\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)$$

$$-\left(\frac{1}{N}\sum_{i=1}^{N}g\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right),X_i\right)-\frac{1}{N}\sum_{i=1}^{N}g\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)\right)$$

$$=\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\hat{g}}{\partial w}\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right)-F_W^{-1}\left(F_X(X_i)\right)\right)$$

$$+\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2\hat{g}}{\partial w^2}\left(\tilde{w}_i,X_i\right)\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right)-F_W^{-1}\left(F_X(X_i)\right)\right)^2 \tag{1.75}$$

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{\partial g}{\partial w}\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right)-F_W^{-1}\left(F_X(X_i)\right)\right)$$

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 g}{\partial w^2}\left(\tilde{w}_i,X_i\right)\cdot\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right)-F_W^{-1}\left(F_X(X_i)\right)\right)^2. \tag{1.76}$$

$$=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial\hat{g}}{\partial w}\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)-\frac{\partial g}{\partial w}\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)\right)$$

$$\times\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right)-F_W^{-1}\left(F_X(X_i)\right)\right)+o_p\left(N^{-1/2}\right).$$

$$\leq\sup_{x\in\mathbb{X}}\left|\frac{\partial\hat{g}}{\partial w}\left(F_W^{-1}\left(F_X(x)\right),x\right)-\frac{\partial g}{\partial w}\left(F_W^{-1}\left(F_X(x)\right),x\right)\right| \tag{1.77}$$

$$\times\sup_{x\in\mathbb{X}}\left|\left(\hat{F}_W^{-1}\left(\hat{F}_X(x)\right)-F_W^{-1}\left(F_X(x)\right)\right)\right|+o_p\left(N^{-1/2}\right). \tag{1.78}$$

In the second to last equality we used the fact that (1.76) is $o_p(N^{-1/2})$ because $\partial^2 g(w,x)/\partial w^2$ is bounded and because $(\hat{F}_W^{-1}(\hat{F}_X(X_i))-F_W^{-1}(F_X(X_i)))^2$ is $o_p(N^{-1/2}$ by Lemma A.6. In addition (1.75) is $o_p(N^{-1/2})$ because $\sup_{w,x}|\partial^2\hat{g}(w,x)/\partial w^2-\partial^2 g(w,x)/\partial w^2|=o_p(1)$, so $\sup_{w,x}|\partial^2\hat{g}(w,x)/\partial w^2|\leq 2\sup_{w,x}|\partial^2 g(w,x)/\partial w^2|$ with probability approaching 1.

By Lemma **??**, $\sup_{x\in\mathbb{X}}|\frac{\partial\hat{g}}{\partial w}(F_W^{-1}(F_X(x)),x)-\frac{\partial g}{\partial w}(F_W^{-1}(F_X(x)),x)|=o_p(N^{-\eta})$. Using Lemma A.6 it follows that $\sup_{x\in\mathbb{X}}|(\hat{F}_W^{-1}(\hat{F}_X(x))-F_W^{-1}(F_X(x)))|=o_p(N^{-(1/2-\eta/2)}$, so that the product in (1.77)-(1.78) is $o_p(N^{-1/2})$. This finishes the part of the proof showing that the sum of (1.67) and (1.68) is $o_p(N^{-1/2})$.

Next, consider the sum of (1.70) and (1.71). First we show that

$$\frac{1}{N}\sum_{i=1}^{N}\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right)-\frac{1}{N}\sum_{i=1}^{N}F_W^{-1}\left(\hat{F}_X(X_i)\right) \tag{1.79}$$

$$-\left(\frac{1}{N}\sum_{i=1}^{N}\hat{F}_W^{-1}\left(F_X(X_i)\right)-\frac{1}{N}\sum_{i=1}^{N}F_W^{-1}\left(F_X(X_i)\right)\right)=o_p\left(N^{-1/2}\right) \tag{1.80}$$

By Lemma A.6

$$\frac{1}{N}\sum_{i=1}^{N}\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right)-\frac{1}{N}\sum_{i=1}^{N}F_W^{-1}\left(\hat{F}_X(X_i)\right)$$

$$-\left(\frac{1}{N}\sum_{i=1}^{N}\hat{F}_W^{-1}\left(F_X(X_i)\right)-\frac{1}{N}\sum_{i=1}^{N}F_W^{-1}\left(F_X(X_i)\right)\right)$$

$$=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{f_W\left(F_W^{-1}\left(\hat{F}_X(X_i)\right)\right)}\cdot\left(\hat{F}_W\left(F_W^{-1}\left(\hat{F}_X(X_i)\right)\right)-\hat{F}_X(X_i)\right) \tag{1.81}$$

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{1}{f_W\left(F_W^{-1}\left(F_X(X_i)\right)\right)}\cdot\left(\hat{F}_W\left(F_W^{-1}\left(F_X(X_i)\right)\right)-F_X(X_i)\right) \tag{1.82}$$

$$+\frac{1}{N}\sum_{i=1}^{N}\frac{1}{f_W\left(F_W^{-1}\left(F_X(X_i)\right)\right)}\cdot\left(\hat{F}_W\left(F_W^{-1}\left(\hat{F}_X(X_i)\right)\right)-\hat{F}_X(X_i)\right) \tag{1.83}$$

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{1}{f_W\left(F_W^{-1}\left(F_X(X_i)\right)\right)}\cdot\left(\hat{F}_W\left(F_W^{-1}\left(\hat{F}_X(X_i)\right)\right)-\hat{F}_X(X_i)\right)+o_p\left(N^{-1/2}\right) \tag{1.84}$$

$$=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{f_W\left(F_W^{-1}\left(\hat{F}_X(X_i)\right)\right)}-\frac{1}{f_W\left(F_W^{-1}\left(\hat{F}_X(X_i)\right)\right)}\right) \tag{1.85}$$

$$\times\left(\hat{F}_W\left(F_W^{-1}\left(\hat{F}_X(X_i)\right)\right)-\hat{F}_X(X_i)\right) \tag{1.86}$$

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{1}{f_W\left(F_W^{-1}\left(F_X(X_i)\right)\right)} \tag{1.87}$$

$$\times\left(\left(\hat{F}_W\left(F_W^{-1}\left(F_X(X_i)\right)\right)-F_X(X_i)\right)-\left(\hat{F}_W\left(F_W^{-1}\left(\hat{F}_X(X_i)\right)\right)-\hat{F}_X(X_i)\right)\right) \tag{1.88}$$

Lemma A.1 implies that $\sup_{q\in[0,1]}|\hat{F}_W(F_W^{-1}(q))-q|=o_p(N^{-\delta}$ for any $\delta<1/2$. The same Lemma, combined with continuous differentiability of $F_W(w)$, implies that $\sup_{x\in\mathbb{X}}|f_W(F_W^{-1}(\hat{F}_X(x)))^{-1}-f_W(F_W^{-1}(F_X(x)))^{-1}|=o_p(N^{-\delta})$. Hence the product (1.85)-(1.86) is $o_p(N^{-1/2})$.

To show that the product (1.87)-(1.88) is $o_p(N^{-1/2})$ it is sufficient to show that

$$\sup\left|\left(\hat{F}_W\left(F_W^{-1}\left(F_X(x)\right)\right)-F_X(x)\right)-\left(\hat{F}_W\left(F_W^{-1}\left(\hat{F}_X(x)\right)\right)-\hat{F}_X(x)\right)\right|=o_p\left(N^{-1/2}\right) \tag{1.89}$$

since we can ignore the factor in (1.85) because it is bounded. We can rewrite this as

$$\sup\left|\left(\hat{F}_W\left(F_W^{-1}\left(F_X(x)\right)\right)-F_X(x)\right)-\left(\hat{F}_W\left(F_W^{-1}\left(\hat{F}_X(x)\right)\right)-\hat{F}_X(x)\right)\right| \tag{1.90}$$

$$\sup\left|\left(\hat{F}_W\left(F_W^{-1}\left(F_X(x)\right)\right)-\hat{F}_W\left(F_W^{-1}\left(\hat{F}_X(x)\right)\right)\right)\right. \tag{1.91}$$

$$-\left(F_W^{-1}\left(F_W\left(F_X(x)\right)\right)-F_W^{-1}\left(F_W\left(\hat{F}_X(x)\right)\right)\right)\Big|, \quad (1.92)$$

which is $o_p(N^{-1/2})$ by an application of Lemma A.4, e.g., with $\eta = 2/3$ and $\delta = 2/5$. This shows that (1.79)-(1.80) holds.

In addition, for all $\delta < 1/2$, by Lemma A.2,

$$\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right)-F_W^{-1}\left(\hat{F}_X(X_i)\right)=o_p\left(N^{-\delta}\right), \quad (1.93)$$

and

$$\hat{F}_W^{-1}\left(F_X(X_i)\right)-F_W^{-1}\left(F_X(X_i)\right)=o_p\left(N^{-\delta}\right). \quad (1.94)$$

In order to prove that the sum (1.70) and (1.71) is $o_p(N^{-1/2})$, it is now sufficient to prove that with a continuously differentiable function $g(w,x)$, for some $\delta < 1/2$ and some $\eta > 0$,

$$\sup_{w\in\mathbb{W},x\in\mathbb{X},|a|,|b|,|c|<N^{-\delta},|-a-b+c|<N^{-1/2-\eta}}|g(w,x)-g(w+a,x)-(g(w+b,x)-g(w+c,x)|=o_p\left(N^{-1/2}\right)$$

Using a Taylor series expansion we can write

$$\begin{aligned}
&g(w,x)-g(w+a,x)-(g(w+b,x)-g(w+c,x)\\
&\quad -\frac{\partial g}{\partial w}(w,x)\cdot a-\frac{1}{2}\frac{\partial^2 g}{\partial w^2}(w_a,x)\cdot a^2-\frac{\partial g}{\partial w}(w,x)\cdot b-\frac{1}{2}\frac{\partial^2 g}{\partial w^2}(w_b,x)\cdot b^2\\
&\qquad +\frac{\partial g}{\partial w}(w,x)\cdot c-\frac{1}{2}\frac{\partial^2 g}{\partial w^2}(w_c,x)\cdot c^2\\
&-(a+b-c)\cdot\frac{\partial g}{\partial w}(w,x)-\frac{1}{2}\frac{\partial^2 g}{\partial w^2}(w_a,x)\cdot a^2-\frac{1}{2}\frac{\partial^2 g}{\partial w^2}(w_b,x)\cdot b^2+\frac{1}{2}\frac{\partial^2 g}{\partial w^2}(w_c,x)\cdot c^2=o_p\left(N^{-1/2}\right).
\end{aligned}$$

□

Lemma A.13

$$\frac{1}{N}\sum_{i=1}^{N}\hat{g}\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)-\frac{1}{N}\sum_{i=1}^{N}g\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)=o_p\left(\right)$$

Proof of Lemma 4.2:
By a mean value theorem there are a_i such that

$$\begin{aligned}
&g\left(\hat{F}_W^{-1}\left(F_X(X_i)\right),X_i\right)\\
&\qquad =g\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)+g_w\left(a_i,X_i\right)\cdot\left(\hat{F}_W^{-1}\left(F_X(X_i)\right)-F_W^{-1}\left(F_X(X_i)\right)\right)
\end{aligned}$$

and $|a_i - F_W^{-1}(F_X(X_i))| \leq |\hat{F}_W^{-1}(F_X(X_i)) - F_W^{-1}(F_X(X_i))|$. Thus

$$\begin{aligned}
&g\left(\hat{F}_W^{-1}\left(F_X(X_i)\right),X_i\right)\\
&\qquad =g\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)+g_w\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)\cdot\left(\hat{F}_W^{-1}\left(F_X(X_i)\right)-F_W^{-1}\left(F_X(X_i)\right)\right)
\end{aligned}$$

$$+ \left(g_w\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) - g_w\left(a_i, X_i\right)\right) \cdot \left(\hat{F}_W^{-1}\left(F_X(X_i)\right) - F_W^{-1}\left(F_X(X_i)\right)\right).$$

Consider the second term on the right-hand side. By a mean value theorem and by Assumption **??** we have for some $b_i \in \mathbb{W}$:

$$\left(g_w\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) - g_w\left(a_i, X_i\right)\right) \cdot \left(\hat{F}_W^{-1}\left(F_X(X_i)\right) - F_W^{-1}\left(F_X(X_i)\right)\right)$$

$$= g_{ww}(b_i, X_i) \cdot \left(F_W^{-1}\left(F_X(X_i)\right) - a_i\right) \cdot \left(\hat{F}_W^{-1}\left(F_X(X_i)\right) - F_W^{-1}\left(F_X(X_i)\right)\right).$$

Because $|a_i - F_W^{-1}(F_X(X_i))| \leq |\hat{F}_W^{-1}(F_X(X_i)) - F_W^{-1}(F_X(X_i))|$ this can be bounded in absolute value by

$$|g_{ww}(b_i, X_i)| \cdot \left(\hat{F}_W^{-1}\left(F_X(X_i)\right) - F_W^{-1}\left(F_X(X_i)\right)\right)^2$$

$$\leq |g_{ww}(b_i, X_i)| \cdot \sup_{q \in [0,1]} \left(\hat{F}_W^{-1}(q) - F_W^{-1}(q)\right)^2$$

Because $\sup_{q \in [0,1]} N^{\delta}(\hat{F}_W^{-1}(q) - F_W^{-1}(q)) = o_p(1)$ for all $\delta < 1/2$, it follows that the last term is $o_p(N^{-2\delta})$ for all $\delta < 1/2$ and therefore $o_p(N^{-1/2})$, and thus

$$g\left(\hat{F}_W^{-1}\left(F_X(X_i)\right), X_i\right) = g\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right)$$

$$+ g_w\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) \cdot \left(\hat{F}_W^{-1}\left(F_X(X_i)\right) - F_W^{-1}\left(F_X(X_i)\right)\right) + o_p\left(N^{-1/2}\right).$$

The remaining step is to show that

$$\hat{\mu}_W - \frac{1}{N}\sum_{i=1}^{N} g_w\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) \cdot \left(\hat{F}_W^{-1}\left(F_X(X_i)\right) - F_W^{-1}\left(F_X(X_i)\right)\right) = o_p\left(N^{-1/2}\right).$$

By Lemma A.10 it is sufficient to show that

$$\hat{\mu}_{WX} - \frac{1}{N}\sum_{i=1}^{N} g_w\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) \cdot \left(\hat{F}_W^{-1}\left(F_X(X_i)\right) - F_W^{-1}\left(F_X(X_i)\right)\right) = o_p\left(N^{-1/2}\right). \quad (1.95)$$

First we work on the second term:

$$\left|\frac{1}{N}\sum_{i=1}^{N} g_w\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) \cdot \left(\hat{F}_W^{-1}\left(F_X(X_i)\right) - F_W^{-1}\left(F_X(X_i)\right)\right)\right.$$

$$\left. - \frac{1}{N}\sum_{i=1}^{N} \frac{g_w\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right)}{f_W(F_W^{-1}(F_X(X_i)))}\left(\hat{F}_W(F_W^{-1}(F_X(X_i))) - F_X(X_i)\right)\right| = o_p\left(N^{-1/2}\right),$$

by Lemma A.6. Substituting in for $\hat{\mu}_{WX}$ we are then left with:

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} q_{WX}(W_i, X_j)$$

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{g_w\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)}{f_W(F_W^{-1}(F_X(X_i)))}\left(\hat{F}_W(F_W^{-1}(F_X(X_i)))-F_X(X_i)\right)$$

$$=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}q_{WX}(W_i,X_j)$$

$$-\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{g_w\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)}{f_W(F_W^{-1}(F_X(X_i)))}\left(1\{W_j\leq F_W^{-1}(F_X(X_i))\}-F_X(X_i)\right)$$

$$=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}q_{WX}(W_i,X_j)$$

$$-\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{g_w\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)}{f_W(F_W^{-1}(F_X(X_i)))}\left(1\{F_W(W_j)\leq F_X(X_i)\}-F_X(X_i)\right),$$

which is equal to zero. □

Proof of Lemma ??: By Lemma A.11 it is sufficient to show that

$$\frac{1}{N}\sum_{i=1}^{N}g\left(F_W^{-1}\left(\hat{F}_X(X_i)\right),X_i\right)-\frac{1}{N}\sum_{i=1}^{N}g\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)-\hat{\mu}_{X_1X_2}=o_p\left(N^{-1/2}\right)$$

Using a Taylor series expansion we obtain:

$$\frac{1}{N}\sum_{i=1}^{N}g\left(F_W^{-1}\left(\hat{F}_X(X_i)\right),X_i\right)-\frac{1}{N}\sum_{i=1}^{N}g\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)$$

$$=\frac{1}{N}\sum_{i=1}^{N}\frac{g_w\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)}{f_W\left(F_W^{-1}\left(F_X(X_i)\right)\right)}\cdot\left(\hat{F}_X(X_i)-F_X(X_i)\right)$$

$$+O_p\left(\left(\hat{F}_X(X_i)-F_X(X_i)\right)^2\right)$$

$$=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{g_w\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)}{f_W\left(F_W^{-1}\left(F_X(X_i)\right)\right)}\cdot\left(1\{X_j\leq X_i\}-F_X(X_i)\right)+o_p\left(N^{-1/2}\right)$$

$$=\hat{\mu}_{X_1X_2}+o_p\left(N^{-1/2}\right)$$

□

Lemma A.14

$$\frac{1}{N}\sum_{i=1}^{N}g\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)-\mathbb{E}\left[g\left(F_W^{-1}\left(F_X(X)\right),X\right)\right]\xrightarrow{d}\mathcal{N}\left(0,\mathbb{E}\left[\left(g\left(F_W^{-1}\left(F_X(X)\right),X\right)-\theta\right)^2\right]\right).$$

Proof: This follows directly from a central limit theorem and the smoothness assumptions on $g(w,x)$, $F_W(w)$ and $F_X(x)$. $\square$.

We are interested in the asymptotic distribution of

$$\hat{\gamma} = \hat{\beta}^{\text{pam}} - \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) \tag{1.96}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\hat{g}\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) - g\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right)\right).$$

Let

$$V = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix},$$

with V_1, V_2, and V vectors of dimension k_1, k_2, and $k = k_1 + k_2$.

Let $r(u)$ be a function from $\mathbb{V}_2$ to $\mathbb{V}_1$, with derivative $r'(u) = \frac{\partial r}{\partial u}(u)$.

We are interested in the distribution of

$$\frac{1}{N}\sum_{i=1}^{N}\left(\hat{g}\left(r(V_{2i}), V_{2i}\right) - g\left(r(V_{2i}), V_{2i}\right)\right),$$

where $\hat{g}(v_1, v_2)$ is a kernel estimator. The kernel is $\mathcal{K}(v)$, and the estimator for $g(v)$ is

$$\hat{g}(v) = \sum_{i=1}^{N} Y_i \cdot \frac{K_\sigma(V_i - v)}{\sum_{j=1}^{N} K_\sigma(V_j - v)},$$

with

$$K_\sigma(v) = \frac{1}{\sigma^k}\cdot\mathcal{K}(v/\sigma),$$

so that

$$K_\sigma(v_1, v_2) = \frac{1}{\sigma^k}\cdot\mathcal{K}(v_1/\sigma, v_2/\sigma).$$

Lemma A.15 (Newey (1994), Lemma B.4)
Let $h(v) = \mathbb{E}[Y|V=v]\cdot f_V(v)$ with $\hat{h}(v)$ a kernel estimator for $h(v)$,

$$\hat{h}(v) = \frac{1}{N}\sum_{i=1}^{N} Y_i \cdot K_\sigma(V_i - v),$$

*with the kernel $K_\sigma(v)$ satisfying the conditions in Assumption **??**. Let the (Y_i, V_i) satisfy the conditions in Assumption **??**, and let b_N satisfy the bandwidth conditions in Assumption 4.11. Then*

$$\sup_v \left|\hat{h}(v) - h(v)\right| = O_p\left(\ln(N)^{1/2}\cdot\left(N b_N^k\right)^{-1/2} + b_N^s\right)$$

Lemma A.16 *Let $k(v)$ be a function from $\mathbb{V} \subset \mathbb{R}^k$ to $\mathbb{V}$, with $\mathbb{V}$ compact, and $k(v)$ continuously differentiable on its domain. Let $g : \mathbb{V} \to \mathbb{R}$ be continuously differentiable, and let $h(v) = \mathbb{E}[Y|V = v] \cdot f_V(v)$ with $\hat{h}(v)$ a kernel estimator for $h(v)$,*

$$\hat{h}(v) = \frac{1}{N} \sum_{i=1}^{N} Y_i \cdot K_\sigma(V_i - v),$$

with the kernel $K_\sigma(v)$ satisfying the conditions in Assumption **??**. *Let the (Y_i, V_i) satisfy the conditions in Assumption* **??**, *and let b_N satisfy the bandwidth conditions in Assumption 4.11. Then*

$$U_1 = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{h}(k(V_i)) - h(k(V_i)) \right) \cdot g(V_i) - \int \left(\hat{h}(k(u)) - h(k(u)) \right) \cdot g(u) f_V(u) du$$

$$= o_p\left(N^{-1/2}\right).$$

Proof of Lemma A.16: There are three steps in the proof. First we obtain a U-statistic representation for U: $U_1 = U_2 + o_p(N^{-1/2})$ with

$$U_2 = \frac{1}{N} \sum_{i=1}^{N} \sum_{j>i} m_2(Y_i, V_i, Y_j, V_j), \tag{1.97}$$

with $m_2(\cdot)$ symmetric: $m_2(y_0, v_0, y_1, v_1) = m_2(y_1, v_1, y_0, v_0)$. Second, using U-statistic results we obtain an asymptotically linear representation for U_2: $U_2 = U_3 + o_p(N^{-1/2})$ with

$$U_3 = \frac{1}{N} \sum_{i=1}^{N} m_3(V_i), \tag{1.98}$$

with $\mathbb{E}[m_3(V)] = 0$. Third, we show that

$$\sup_{v \in \mathbb{V}} m_3(v) \to 0.$$

This implies that $U_3 = o_p(N^{-1/2})$.

□

Proof of Lemma ??: There are two steps in the proof. First we linearize the estimator and second we apply Lemma A.16. First define

$$h_1(v) = f_V(v),$$

and

$$h_2(v) = \mathbb{E}[Y|V = v] \cdot f_V(v),$$

with corresponding kernel estimators

$$\hat{h}_1(v) = \frac{1}{N} \sum_{i=1}^{N} K_\sigma(V_i - v),$$

and

$$\hat{h}_2(v) = \frac{1}{N}\sum_{i=1}^{N} Y_i \cdot K_\sigma(V_i - v).$$

Then

$$g(v) = \frac{h_2(v)}{h_1(v)}, \qquad \text{and} \;\; \hat{g}(v) = \frac{\hat{h}_2(v)}{\hat{h}_1(v)}.$$

The first step is to show that we can linearize the difference between the estimator and the estimand:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\hat{h}_2(r(V_{i2}),V_{i2})}{\hat{h}_1(r(V_{i2}),V_{i2})} - \frac{1}{N}\sum_{i=1}^{N}\frac{h_2(r(V_{i2}),V_{i2})}{h_1(r(V_{i2}),V_{i2})}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{\hat{h}_2(r(V_{i2}),V_{i2}) - h_2(r(V_{i2}),V_{i2})}{h_1(r(V_{i2}),V_{i2})}$$

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{h_2(r(V_{i2}),V_{i2})}{h_1^2(r(V_{i2}),V_{i2})}\hat{h}_1(r(V_{i2}),V_{i2}) - h_1(r(V_{i2}),V_{i2}) + o_p\left(N^{-1/2}\right).$$

The linearization remainder term is:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\hat{h}_2(r(V_{i2}),V_{i2})}{\hat{h}_1(r(V_{i2}),V_{i2})} - \frac{1}{N}\sum_{i=1}^{N}\frac{h_2(r(V_{i2}),V_{i2})}{h_1(r(V_{i2}),V_{i2})}$$

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{\hat{h}_2(r(V_{i2}),V_{i2}) - h_2(r(V_{i2}),V_{i2})}{h_1(r(V_{i2}),V_{i2})}$$

$$+\frac{1}{N}\sum_{i=1}^{N}\frac{h_2(r(V_{i2}),V_{i2})}{h_1^2(r(V_{i2}),V_{i2})}\hat{h}_1(r(V_{i2}),V_{i2}) - h_1(r(V_{i2}),V_{i2})$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{\hat{h}_2(r(V_{i2}),V_{i2})}{\hat{h}_1(r(V_{i2}),V_{i2})} - \frac{1}{N}\sum_{i=1}^{N}\frac{h_2(r(V_{i2}),V_{i2})}{\hat{h}_1(r(V_{i2}),V_{i2})}$$

$$+\frac{1}{N}\sum_{i=1}^{N}\frac{h_2(r(V_{i2}),V_{i2})}{\hat{h}_1(r(V_{i2}),V_{i2})} - \frac{1}{N}\sum_{i=1}^{N}\frac{h_2(r(V_{i2}),V_{i2})}{h_1(r(V_{i2}),V_{i2})}$$

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{\hat{h}_2(r(V_{i2}),V_{i2}) - h_2(r(V_{i2}),V_{i2})}{h_1(r(V_{i2}),V_{i2})}$$

$$+\frac{1}{N}\sum_{i=1}^{N}\frac{h_2(r(V_{i2}),V_{i2})}{h_1^2(r(V_{i2}),V_{i2})}\hat{h}_1(r(V_{i2}),V_{i2}) - h_1(r(V_{i2}),V_{i2})$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{\hat{h}_2(r(V_{i2}),V_{i2}) - h_2(r(V_{i2}),V_{i2})}{\hat{h}_1(r(V_{i2}),V_{i2})} \tag{1.99}$$

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{\hat{h}_2(r(V_{i2}),V_{i2})-h_2(r(V_{i2}),V_{i2})}{h_1(r(V_{i2}),V_{i2})} \tag{1.100}$$

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{h_2(r(V_{i2}),V_{i2})}{\hat{h}_1(r(V_{i2}),V_{i2})h_1(r(V_{i2}),V_{i2})}\cdot\left(\hat{h}_1(r(V_{i2}),V_{i2})-h_1(r(V_{i2}),V_{i2})\right) \tag{1.101}$$

$$+\frac{1}{N}\sum_{i=1}^{N}\frac{h_2(r(V_{i2}),V_{i2})}{h_1^2(r(V_{i2}),V_{i2})}\cdot\left(\hat{h}_1(r(V_{i2}),V_{i2})-h_1(r(V_{i2}),V_{i2})\right) \tag{1.102}$$

By Lemma A.15 $\sup_v|\hat{h}_j(v)-h_j(v)|=O_p(\ln(N)^{1/2}\cdot\left(Nb_N^k\right)^{-1/2}+b_N^s)$. By the rate conditions on b_H, this is $o_p(N^{-1/4})$. By Assumption **??** it follows that $\sup_v|1/\hat{h}_j(v)-1/h_j(v)|$ is also $o_p(N^{-1/4})$. This implies that the sum of (1.122) and (1.123) and the sum of (1.101) and (1.102) are both $o_p\left(N^{-1/2}\right)$. This completes the first step.

The second step applies Lemma A.16 twice. First with $k(v)=(r(v_2)',v_2')'$, $h(v)=h_1(v)$, $g(v)=h_2(r(v_2),v_2)/h_1^2(r(v_2),v_2)$ □

Proof of Lemma ??: The $\psi_{Ni}-\mathbb{E}[\psi_{Ni}]$ for a triangular array with the terms $\psi_{Ni}-\mathbb{E}[\psi_{Ni}]$ and $\psi_{Nj}-\mathbb{E}[\psi_{Nj}]$ and identically distributed. We will apply Liapounov's central limit theorem to this triangular array. First we show that

$$\mathbb{E}\left[\psi_{Ni}\right]\xrightarrow{d}0. \tag{1.103}$$

Define

$$h_1(v_1,v_2)=f_{V_1,V_2}(v_1,v_2)$$

$$h_2(v_1,v_2)=\mathbb{E}[Y|V_1=v_1,V_2=v]\cdot f_{V_1,V_2}(v_1,v_2)$$

with kernel estimators

$$\hat{h}_1(v_1,v_2)=\frac{1}{N}\sum_{i=1}^{N}K_{\sigma_N}\left(V_{1i}-v_1,V_{2i}-v_2\right)$$

$$\hat{h}_2(v_1,v_2)=\frac{1}{N}\sum_{i=1}^{N}Y_i\cdot K_{\sigma_N}\left(V_{1i}-v_1,V_{2i}-v_2\right)$$

Then

$$\begin{aligned}\mathbb{E}\left[\psi_{Ni}\right]&=\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\psi_{Ni}\right]\\&=\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\sigma_N^{k_1/2}\cdot\int(Y_i-g(k(u),u))\cdot K_{\sigma_N}(r(u)-V_{i1},u-V_{2i})du\right]\\&=\sigma_N^{k_1/2}\cdot\mathbb{E}\left[\int\frac{1}{N}\sum_{i=1}^{N}(Y_i-g(k(u),u))\cdot K_{\sigma_N}(r(u)-V_{i1},u-V_{2i})du\right]\\&=\sigma_N^{k_1/2}\cdot\mathbb{E}\left[\int\left(\hat{h}_2(r(u),u)-\hat{h}_1(r(u),u)\cdot g(r(u),u)\right)du\right]\end{aligned}$$

$$= \sigma_N^{k_1/2} \cdot \int \left(\mathbb{E}\left[\hat{h}_2(r(u),u)\right] - \mathbb{E}\left[\hat{h}_1(r(u),u)\right] \cdot g(r(u),u) \right) du$$

We can bound this by

$$\left| \sigma_N^{k_1/2} \cdot \int \left(\mathbb{E}\left[\hat{h}_2(r(u),u)\right] - \mathbb{E}\left[\hat{h}_1(r(u),u)\right] \cdot g(r(u),u) \right) du \right|$$

$$\leq \left| \sigma_N^{k_1/2} \cdot \int \left(h_2(r(u),u) - h_1(r(u),u) \cdot g(r(u),u) \right) du \right| \tag{1.104}$$

$$+ \sigma_N^{k_1/2} \cdot \int \sup_{v_1,v_2} \left| \mathbb{E}\left[\hat{h}_2(v_1,v_2)\right] - h_2(v_1,v_2) \right| du \tag{1.105}$$

$$+ \sigma_N^{k_1/2} \cdot \int \sup_{v_1,v_2} \left| \mathbb{E}\left[\hat{h}_1(v_1,v_2)\right] - h_1(v_1,v_2) \right| \cdot |g(r(u),u)| \, du \tag{1.106}$$

(1.104) is equal to $C \cdot \sigma_N^{k_1/2}$ which converges to zero. By uniform convergence of $\mathbb{E}[\hat{h}_1(v_1,v_2)$ and $\mathbb{E}[\hat{h}_2(v_1,v_2)$ to $h_1(v_1,v_2)$ and $h_2(v_1,v_2)$ respectively it follows that (1.105) and (1.106) converge to zero. This demonstrates (1.103).

Second, we show that

$$\mathbb{E}\left[(\psi_{Ni})^2\right] \xrightarrow{d} \Omega. \tag{1.107}$$

First,

$$\mathbb{E}\left[(\psi_{Ni})^2\right] = \sigma_N^{k_1} \cdot \mathbb{E}\left[\left(\int \left(Y_i - g(r(u),u) \right) K_{\sigma_N}(r(u) - V_{i1}, u - V_{i2}) du \right)^2 \right]$$

$$= \sigma_N^{k_1} \cdot \mathbb{E}\left[\left(\int \left(g(V_{i1},V_{i2}) - g(r(u),u) \right) K_{\sigma_N}(r(u) - V_{i1}, u - V_{i2}) du \right)^2 \right] \tag{1.108}$$

$$+ \sigma_N^{k_1} \cdot \mathbb{E}\left[\sigma^2(V_{i1},V_{i2}) \left(\int K_{\sigma_N}(r(u) - V_{i1}, u - V_{i2}) du \right)^2 \right]. \tag{1.109}$$

We show that (1.108) is $o_p(1)$ and that (1.109) converges to Ω. First, (1.108):

$$\sigma_N^{k_1} \cdot \mathbb{E}\left[\left(\int \left(g(V_{i1},V_{i2}) - g(r(u),u) \right) K_{\sigma_N}(r(u) - V_{i1}, u - V_{i2}) du \right)^2 \right]$$

$$= \frac{1}{\sigma_N^{k_1+2k_2}} \cdot \mathbb{E}\left[\left(\int \left(g(V_{i1},V_{i2}) - g(r(u),u) \right) \mathcal{K}\left(\frac{r(u) - V_{i1}}{\sigma}, \frac{u - V_{i2}}{\sigma} \right) du \right)^2 \right]$$

$$= \frac{1}{\sigma_N^{k_1+2k_2}} \cdot \int \left(\int \left(g(v_1,v_2) - g(r(u),u) \right) \mathcal{K}\left(\frac{r(u) - v_1}{\sigma}, \frac{u - v_2}{\sigma} \right) du \right)^2 f_{V_1,V_2}(v_1,v_2) dv_1 dv_2.$$

By change of variables from v_1 to $s = (r(v_2) - v_1)/\sigma$, with Jacobian $|J| = \sigma^{k_1}$ this is equal to

$$\frac{1}{\sigma_N^{2k_2}} \cdot \int \left(\int \left(g(r(v_2 - s\sigma, v_2) - g(r(u),u) \right) \mathcal{K}\left(\frac{r(u) - r(v_2)}{\sigma} + s, \frac{u - v_2}{\sigma} \right) du \right)^2 f_{V_1,V_2}(r(v_1) - s\sigma, v_2) ds dv_2.$$

By change of variable in the inner integral from u to $t = (u - v_2)/\sigma$ with Jacobian $|J| = \sigma^{k_2}$ this is equal to

$$\int \left(\int \left(g(r(v_2 - s\sigma, v_2) - g(r(v_2 + t\sigma), v_2 + t\sigma)\right) \mathcal{K} \left(\frac{r(v_2 + t\sigma) - r(v_2)}{\sigma} + s, t \right) dt \right)^2$$

$$\times f_{V_1,V_2}(r(v_2) - s\sigma, v_2) ds dv_2. \tag{1.110}$$

There is are compact sets $\mathbb{S}$ and $\mathbb{T}$ such that

$$\mathcal{K} \left(\frac{r(v_2 + t\sigma) - r(v_2)}{\sigma} + s, t \right)$$

is zero for $(s, t) \notin \mathbb{S} \times \mathbb{T}$. For $(s, t) \in \mathbb{S} \times \mathbb{T}$ we can bound

$$|g(r(v_2 - s\sigma, v_2) - g(r(v_2 + t\sigma), v_2 + t\sigma)| \leq C\sigma$$

for some constant C. Hence we can bound (1.110) by

$$C\sigma \int 1\{s \in \mathbb{S}\} \left(\int 1\{t \in \mathbb{T}\} \mathcal{K} \left(\frac{r(v_2 + t\sigma) - r(v_2)}{\sigma} + s, t \right) dt \right)^2 f_{V_1,V_2}(r(v_2) - s\sigma, v_2) ds dv_2 \leq C_2 \sigma.$$

where the last inequality follows from the boundedness of $\mathcal{K}(v_1, v_2)$ and $f_{V_1,V_2}(v_1, v_2)$ and the compactness of $\mathbb{S}$, $\mathbb{T}$ and $\mathbb{V}_2$. This shows that (1.108) is $o_p(1)$. Next, consider (1.109): First,

$$\sigma_N^{k_1} \cdot \mathbb{E} \left[\sigma^2(V_1, V_2) \left(\int K_{\sigma_N}(r(u) - V_1, u - V_2) du \right)^2 \right]$$

$$= \frac{1}{\sigma_N^{k_1 + 2k_2}} \cdot \mathbb{E} \left[\sigma^2(V_1, V_2) \left(\int \mathcal{K} \left(\frac{r(u) - V_1}{\sigma}, \frac{u - V_2}{\sigma} \right) du \right)^2 \right]$$

$$= \frac{1}{\sigma_N^{k_1 + 2k_2}} \cdot \int \sigma^2(v_1, v_2) \left(\int \mathcal{K} \left(\frac{r(u) - v_1}{\sigma}, \frac{u - v_2}{\sigma} \right) du \right)^2 f_{V_1,V_2}(v_1, v_2) dv_1 dv_2.$$

By change of variable from v_1 to $s = (r(v_2) - v_1)/\sigma$ with Jacobian $|J| = \sigma_N^{k_1}$ this is equal to:

$$\frac{1}{\sigma_N^{2k_2}} \cdot \int \sigma^2(r(v_1) - s\sigma, v_2) \left(\int \mathcal{K} \left(\frac{r(u) - r(v_1)}{\sigma} + s, \frac{u - v_2}{\sigma} \right) du \right)^2 f_{V_1,V_2}(r(v_2) - s\sigma, v_2) ds dv_2.$$

By change of variables from u to $t = (u - v_2)/\sigma_N$ with Jacobian $|J| = \sigma_N^{k_2}$ this is equal to

$$\frac{1}{\sigma_N^{2k_2}} \cdot \int \sigma^2(r(v_1) - s\sigma, v_2) \left(\int \sigma_N^{k_2} \cdot \mathcal{K} \left(\frac{r(v_2 + t\sigma) - r(v_1)}{\sigma} + s, t \right) dt \right)^2 f_{V_1,V_2}(r(v_2) - s\sigma, v_2) ds dv_2$$

$$= \int \sigma^2(r(v_1) - s\sigma, v_2) \left(\int \mathcal{K} \left(\frac{r(v_2 + t\sigma) - r(v_1)}{\sigma} + s, t \right) du \right)^2 f_{V_1,V_2}(r(v_2) - s\sigma, v_2) ds dv_2. \tag{1.111}$$

To show that (1.111) converges to Ω, we write

$$\sigma^2(r(v_2)-s\sigma,v_2)\left(\int\mathcal{K}\left(\frac{r(v_2+t\sigma)-r(v_1)}{\sigma}+s,t\right)du\right)^2 f_{V_1,V_2}(r(v_2)-s\sigma,v_2)dv_1dv_2-\Omega$$

$$=\left(\sigma^2(r(v_2)-s\sigma,v_2)\left(\int\mathcal{K}\left(\frac{r(v_2+t\sigma)-r(v_1)}{\sigma}+s,t\right)du\right)^2 f_{V_1,V_2}(r(v_2)-s\sigma,v_2)dv_1dv_2\right. \tag{1.112}$$

$$\left.-\sigma^2(r(v_2),v_2)\left(\int\mathcal{K}\left(\frac{r(v_2+t\sigma)-r(v_1)}{\sigma}+s,t\right)du\right)^2 f_{V_1,V_2}(r(v_2)-s\sigma,v_2)dv_1dv_2\right) \tag{1.113}$$

$$+\left(\sigma^2(r(v_2),v_2)\left(\int\mathcal{K}\left(\frac{r(v_2+t\sigma)-r(v_1)}{\sigma}+s,t\right)du\right)^2 f_{V_1,V_2}(r(v_2)-s\sigma,v_2)dv_1dv_2\right. \tag{1.114}$$

$$\left.-\sigma^2(r(v_2),v_2)\left(\int\mathcal{K}\left(r'(v_2)t+s,t\right)du\right)^2 f_{V_1,V_2}(r(v_2)-s\sigma,v_2)dv_1dv_2\right) \tag{1.115}$$

$$\left(\sigma^2(r(v_2),v_2)\left(\int\mathcal{K}\left(r'(v_2)t+s,t\right)du\right)^2 f_{V_1,V_2}(r(v_2)-s\sigma,v_2)dv_1dv_2\right. \tag{1.116}$$

$$\left.\sigma^2(r(v_2),v_2)\left(\int\mathcal{K}\left(r'(v_2)t+s,t\right)du\right)^2 f_{V_1,V_2}(r(v_2),v_2)dv_1dv_2\right) \tag{1.117}$$

$$+\int\sigma^2(r(v_2),v_2)\left(\int\mathcal{K}\left(r'(v_2)t+s,t\right)du\right)^2 f_{V_1,V_2}(r(v_2),v_2)dv_1dv_2-\Omega \tag{1.118}$$

We show that (1.112)-(1.113), (1.114)-(1.115) and (1.116)-(1.117)are $o_p(1)$. Combined with the fact that (1.118) is zero by definition, it follows that (1.111) converges to Ω. There are compact sets $\mathbb{S}$ and $\mathbb{T}$ such that for $(s,t)\notin\mathbb{S}\times\mathbb{T}$ the kernel $\mathcal{K}\left((r(v_2+t\sigma)-r(v_1))/\sigma+s,t\right)$ is equal to zero for all values of σ and v_2. For $s\in\mathbb{S}$, $t\in\mathbb{T}$, and $v_2\in\mathbb{V}_2$, $\mathcal{K}\left((r(v_2+t\sigma)-r(v_1))/\sigma+s,t\right)$ and $f_{V_1,V_2}(r(v_2)-s\sigma,v_2)$ are bounded. In addition, for these values $\left|\sigma^2(r(v_1)-s\sigma,v_2)-\sigma^2(r(v_2),v_2)\right|\leq C\sigma$ for some positive constant C. This implies that (1.112)-(1.113) is $o_p(1)$. Similarly, the difference $|(r(v_2+t\sigma)-r(v_1))/\sigma-r'(v_2)t|\leq C\sigma$ for some positive constant C, and hence (1.114)-(1.115) is $o_p(1)$ Finally, for $s\in\mathbb{S}$ and $v_2\in\mathbb{V}_2$ $|f_{V_1,V_2}(r(v_2)-s\sigma,v_2)-f_{V_1,V_2}(r(v_2),v_2)|\leq C\sigma$ for some constant C, which combined with the bounds on $\mathcal{K}\left((r(v_2+t\sigma)-r(v_1))/\sigma+s,t\right)$ and $\sigma(r(v_2),v_2)$ implies that (1.116)-(1.117) is $o_p(1)$.

The combination of (1.103) and (1.107) then implies

$$\mathrm{Var}(\psi_{Ni})\xrightarrow{d}\Omega. \tag{1.119}$$

Third, we show that

$$\mathbb{E}\left[|\psi_{Ni}|^4\right]/N \xrightarrow{d} 0. \tag{1.120}$$

First consider

$$\sigma_N^{k_1} \cdot \int K_{\sigma_N}(r(u)-v_1, u-v_2)du = \sigma_N^{k_1} \cdot \int \frac{1}{\sigma_N^k} K\left(\frac{(r(u)-v_1}{\sigma_N}, \frac{u-v_2}{\sigma_N}\right) du.$$

By a change of variables from u to $t=(u-v_2)/\sigma_N$ with Jacobian $|J|=\sigma_N^{k_2}$ this is equal to

$$\sigma_N^{k_1} \int \frac{1}{\sigma_N^{k_1}} K\left(\frac{(r(t\sigma_N+v_2)-v_1}{\sigma_N}, t\right) dt = \int K\left(\frac{(r(t\sigma_N+v_2)-v_1}{\sigma_N}, t\right) dt.$$

Since $K(v)$ is bounded and equal to zero outside a compact set, it follows that

$$\sup_{v1,v2} \sigma_N^{k_1} \cdot \int K_{\sigma_N}(r(u)-v_1, u-v_2)du \le C, \tag{1.121}$$

for some constant C. Next, consider

$$\begin{aligned}
&\frac{1}{N}\mathbb{E}\left[\left|\psi_{Ni}\right|^4 \middle| V_i = v\right] \\
&\quad = \frac{\sigma_N^{2k_1}}{N} \cdot \mathbb{E}\left[\left(\int (Y_i - g(k(u),u)) K_{\sigma_N}\left(r(u)-V_{i1}, u-V_{i2}\right) du\right)^4 \middle| V_i = v\right].
\end{aligned}$$

By Assumption **??** this can be bounded by

$$\begin{aligned}
&\frac{\sigma_N^{2k_1}}{N} \cdot C \cdot \mathbb{E}\left[\left(\int K_{\sigma_N}\left(r(u)-V_{i1}, u-V_{i2}\right) du\right)^4 \middle| V_i = v\right]. \\
&\quad = \frac{\sigma_N^{2k_1}}{N} \cdot C \cdot \left(\int K\left(r(u)-v_1, u-v_2\right) du\right)^4. \\
&\quad = \frac{1}{\sigma_N^{2k_1} \cdot N} \cdot C \cdot \sigma_N^{4k_1} \left(\int K_{\sigma_N}\left(r(u)-v_1, u-v_2\right) du\right)^4 \le \frac{C_1}{\sigma_N^{2k_1} \cdot N},
\end{aligned}$$

where the last inequality follows from (1.121). This shows that

$$\mathbb{E}\left[|\psi_{Ni}|^4 \middle| V = v\right]/N \le C_1 \cdot \left(N\sigma_N^{2k_1}\right)^{-1},$$

uniformly in v, which in turn implies that (1.120) holds because by Assumption **??**, $(N\sigma_N^{2k_1})^{-1} \to 0$.

The fourth step involving using (1.119) and (1.120) to show that the Liapounov condition

$$\left(\sum_{i=1}^N \mathbb{E}\left[|\psi_{Ni} - \mathbb{E}[\psi_{Ni}]|^{2+\delta}\right]\right)^2 = o\left(\left(\sum_{i=1}^N \mathbb{E}\left[|\psi_{Ni} - \mathbb{E}[\psi_{Ni}]|^2\right]\right)^{2+\delta}\right), \tag{1.122}$$

is satisfied for $\delta = 2$. To see this, note that the Liaponouv condition (1.122) with $\delta = 2$ is equivalent to

$$\sum_{i=1}^{N} \mathbb{E}\left[|\psi_{Ni} - \mathbb{E}[\psi_{Ni}]|^4\right] = o\left(\left(\sum_{i=1}^{N} \mathbb{E}\left[|\psi_{Ni} - \mathbb{E}[\psi_{Ni}]|^2\right]\right)^2\right).$$

By the fact that the ψ_{Ni} have identical distributions for all i for given N this is equivalent to

$$N\mathbb{E}\left[|\psi_{Ni} - \mathbb{E}[\psi_{Ni}]|^4\right] = o\left(N^2\left(\mathbb{E}\left[|\psi_{Ni} - \mathbb{E}[\psi_{Ni}]|^2\right]\right)^2\right),$$

or

$$\mathbb{E}\left[|\psi_{Ni} - \mathbb{E}[\psi_{Ni}]|^4\right]/N = o\left(\left(\mathbb{E}\left[|\psi_{Ni} - \mathbb{E}[\psi_{Ni}]|^2\right]\right)^2\right). \tag{1.123}$$

By (1.119) it follows that $\mathbb{E}\left[|\psi_{Ni} - \mathbb{E}[\psi_{Ni}]|^2\right]$ converges to V and therefore sufficient for (1.123) is

$$\mathbb{E}\left[|\psi_{Ni} - \mathbb{E}[\psi_{Ni}]|^4\right]/N = o(1).$$

This follows from (1.120). □

Theorem A.1 LIAPOUNOV'S CENTRAL LIMIT THEOREM
Let ψ_{Ni}, $i = 1, \ldots, N$, $N = 1, 2, \ldots$ form a triangular array of independent random variables with $\sigma_{Ni}^2 = \mathrm{Var}(\psi_{Ni})$, and let $s_N^2 = \sum_{i=1}^N \sigma_{Ni}^2$. If for some $\delta > 0$,

$$\left(\sum_{i=1}^{N} \mathbb{E}\left[|\psi_{Ni} - \mathbb{E}[\psi_{Ni}]|^{2+\delta}\right]\right)^2 = o\left(\left(s_N^2\right)^{2+\delta}\right), \tag{1.124}$$

then

$$\frac{\frac{1}{\sqrt{N}}\sum_{i=1}^N(\psi_{Ni} - \mathbb{E}[\psi_{Ni}])}{\sqrt{s_N^2/N}} \xrightarrow{d} \mathcal{N}(0,1).$$

For this version of the theorem see Billingsley (1985), Theorem 27.3, page 371.

$$\omega(u) = \begin{pmatrix} -g(r(u),u) \\ 1 \end{pmatrix}$$

$$V = \int \omega(x_2)\int_{\mathcal{X}} K\left(\frac{\partial \overline{x}_1(x_2)}{\partial x_2'}v + u, v\right)\mathrm{d}v\Sigma(\overline{x}_1(x_2), x_2)\omega(x_2)'\int_{\mathcal{X}} K\left(\frac{\partial \overline{x}_1(x_2)}{\partial x_2'}v + u, v\right)\mathrm{d}v\mathrm{d}u\mathrm{d}x_2$$

A.4 Proof of Lemma 5.1

The following Lemma will be useful in the proof to Lemma 5.1 given below.

Lemma A.17 *Let X and $h(q,x)$ respectively be a continuous random variable with finite support $X \in [a,b]$, and a continuously differentiable function in x (over the support of X). The slope coefficient of the (mean squared error minimizing) conditional linear predictor (CLP) of $h(Q,X)$ given X and conditional on $Q=q$ has a weighted average derivative representation of*

$$b(q) = \frac{Cov\left(h(Q,X), X|Q=q\right)}{Var\left(X|Q=q\right)} = \mathbb{E}_{X|W}\left[\omega(Q,X)\frac{\partial h(Q,X)}{\partial x}|Q=q\right],$$

where

$$\omega(s,t) = \frac{1}{dF_{X|Q}(t|s)}\frac{\mathbb{E}_{X|Q}[X-\mu_{X|Q}(Q)|Q=s, X\geq t]\left(1-F_{X|Q}(t|s)\right)}{\int_{r=a}^{r=b}\mathbb{E}_{X|Q}[X-\mu_{X|Q}(Q)|Q=s, X\geq r]\left(1-F_{X|Q}(v|s)\right)dr}$$

with

$$\mathbb{E}_{X|Q}\left[\omega(Q,X)|Q=q\right] = 1,$$

and $\omega(q,x)$ such that $b(q)$ gives maximum weight to values of $\frac{\partial h(Q,X)}{\partial x}$ with X close to its conditional mean, $\mu_{X|Q}(q) = \mathbb{E}_{X|Q}[X|Q=q]$, and minimum weight to values of $\frac{\partial h(Q,X)}{\partial x}$ with X near the boundaries of its support.

Proof: We first derive the weighted average derivative representation of the conditional linear predictor (CLP) slope coefficient. We then use a simple integration by parts argument to characterize the nature of the derivative weights. Observe that $h(Q,X) - h(Q,a) = \int_{t=a}^{t=X}\frac{\partial h(Q,t)}{\partial t}dt$ and that $E_{X|Q}[h(Q,a)(X-\mu_{X|Q}(Q))|Q] = 0$. Under weak conditions we therefore have

$$\begin{aligned}
Cov\left(h(Q,X), X|Q\right) &= \mathbb{E}_{X|Q}[h(Q,X)(X-\mu_{X|Q}(Q))|Q] \\
&= \mathbb{E}_{X|Q}\left[\int_{t=a}^{t=X}\frac{\partial h(Q,t)}{\partial t}(X-\mu_{X|Q}(Q))\mathrm{d}t|Q\right] \\
&= \mathbb{E}_{X|Q}\left[\int_{t=a}^{t=b}\frac{\partial h(Q,t)}{\partial t}(X\geq t)(X-\mu_{X|Q}(Q))\mathrm{d}t|Q\right] \\
&= \int_{t=a}^{t=b}\frac{\partial h(Q,t)}{\partial t}\mathbb{E}_{X|Q}[(X\geq t)(X-\mu_{X|Q}(Q))|Q]\mathrm{d}t \\
&= \int_{t=a}^{t=b}\frac{\partial h(Q,t)}{\partial t}\mathbb{E}_{X|Q}[X-\mu_{X|Q}(Q)|Q, X\geq t]\left(1-F_{X|Q}(t|Q)\right)\mathrm{d}t.
\end{aligned}$$

Similarly, the conditional variance of X can be written as

$$\begin{aligned}
Var\left(X|Q\right) &= \mathbb{E}_{X|Q}[X(X-\mu_{X|Q}(Q))|Q] \\
&= \mathbb{E}_{X|Q}\left[\int_{r=a}^{r=X}1(X-\mu_{X|Q}(Q))\mathrm{d}r|Q\right] \\
&= \int_{r=a}^{r=b}\mathbb{E}_{X|Q}[X-\mu_{X|Q}(Q)|Q, X\geq r]\left(1-F_{X|Q}(v|Q)\right)\mathrm{d}v.
\end{aligned}$$

By the definition of a conditional linear predictor (CLP) we have the required weighted average derivative representation

$$b(q) = \mathbb{E}_{X|Q}\left[\omega(Q,X)\frac{\partial h(Q,X)}{\partial x}|Q=q\right]$$

for $\omega(Q,X)$ as given in the Lemma. That $\mathbb{E}_{X|Q}[\omega(Q,X)|Q=q]=1$ follows immediately. To show that the weighted average derivative representation of $b(Q)$ gives the most emphasis to values of $\frac{\partial h(Q,X)}{\partial t}$ for X close to its conditional mean, begin by noting that

$$\mathbb{E}_{X|Q}\left[\omega(Q,X)\frac{\partial h(Q,X)}{\partial x}|Q\right] = \frac{\int_{t=a}^{t=b}\frac{\partial h(Q,t)}{\partial t}\mathbb{E}_{X|Q}[X-\mu_{X|Q}(Q)|Q,X\geq t]\left(1-F_{X|Q}(t|Q)\right)\mathrm{d}t}{\int_{r=a}^{r=b}\mathbb{E}_{X|Q}[X-\mu_{X|Q}(Q)|Q,X\geq r]\left(1-F_{X|Q}(r|Q)\right)\mathrm{d}r},$$

and hence the weight $b(Q)$ places on $\frac{\partial h(Q,t)}{\partial t}$ is proportional to

$$\mathbb{E}_{X|Q}\left[X-\mu_{X|Q}(Q)|Q,X\geq t\right]\left(1-F_{X|Q}(t|Q)\right). \tag{1.125}$$

Differentiating this expression with respect to t gives

$$\begin{aligned}\frac{\partial}{\partial t}\left\{\mathbb{E}_{X|Q}[X-\mu_{X|Q}(Q)|Q,X\geq t]\left(1-F_{X|Q}(t|Q)\right)\right\} &= \frac{\partial}{\partial t}\int_{s=t}^{s=b} s\ \mathrm{d}F_{X|Q}(s|Q)\\ &\quad -\frac{\partial}{\partial t}\left[1-F_{X|Q}(t|Q)\right]\mu_{X|Q}(Q)\\ &= \frac{\partial}{\partial t}\int_{t}^{b} s\ \mathrm{d}F_{X|Q}(s|Q)\\ &\quad +\mu_{X|Q}(Q)\cdot\mathrm{d}F_{X|Q}(t|Q)\end{aligned}$$

Integration by parts (with $u(s)=1-F_{X|Q}(s|Q)$ and $v(s)=s$) gives

$$\begin{aligned}\int_{s=t}^{s=b}\left[1-F_{X|Q}(s|Q)\right]\ ds &= \left\{1-F_{X|Q}(s|Q)\,s\right\}|_{s=t}^{s=b}+\int_{s=t}^{s=b} s\ dF_{X|Q}(s|Q)\\ &= -\left[1-F_{X|Q}(t|Q)\right]t+\int_{s=t}^{s=b} s\ dF_{X|Q}(s|Q).\end{aligned} \tag{1.126}$$

Using (1.126) we can express the derivative of $\int_t^b s\ \mathrm{d}F_{X|Q}(s|Q)$ as

$$\begin{aligned}\frac{\partial}{\partial t}\int_{t}^{b} s\ dF_{X|Q}(s|Q) &= \frac{\partial}{\partial t}\left\{\left[1-F_{X|Q}(t|Q)\right]t+\int_{s=t}^{s=b}\left[1-F_{X|Q}(s|Q)\right]\ \mathrm{d}s\right\}\\ &= \left[1-F_{X|Q}(t|Q)\right]-t\cdot\mathrm{d}F_{X|Q}(t|Q)+\frac{\partial}{\partial t}\int_{s=t}^{s=b}\left[1-F_{X|Q}(s|Q)\right]\mathrm{d}s\\ &= \left[1-F_{X|Q}(t|Q)\right]-t\cdot\mathrm{d}F_{X|Q}(t|Q)-\left[1-F_{X|Q}(t|Q)\right]\\ &= -t\cdot\mathrm{d}F_{X|Q}(t|Q).\end{aligned}$$

Substituting this result into in the derivative of (1.125) then gives

$$\begin{aligned}&\frac{\partial}{\partial t}\left\{\mathbb{E}_{X|Q}[X-\mu_{X|Q}(Q)|Q,X\geq t]\left(1-F_{X|Q}(t|Q)\right)\right\}\\ &\qquad = -(t-\mu_{X|Q}(Q))\cdot\mathrm{d}F_{X|Q}(t|Q),\end{aligned}$$

which is equal to zero at $t = \mu_{X|Q}(Q)$, is negative for $t > \mu_{X|Q}(Q)$, and positive for $t < \mu_{X|Q}(Q)$, hence the weights attain a maximum at $t = \mu_{X|Q}(Q)$ and minimums at the boundaries of the support of X as claimed.

With the proof of Lemma A.17 completed we now prove Lemma 5.1. We first derive the covariance representation (5.22) of γ. We then derive the weighted average derivative representation (5.23).

Observe that $\beta^{\mathrm{lr}}(\lambda)$ equals

$$\beta^{\mathrm{lr}}(\lambda) = \int_w \int_x \int_z g(F^{-1}_{W|Z}(F_{W_\lambda|Z}(\lambda x + (1-\lambda^2)^{1/2} w|z)|z), x, z) \mathrm{d}F_{W,X,Z}(w, x, z)$$

and hence, by the Chain Rule and the fact that $\mathrm{d}F_{W|Z}(w|z) > 0$ for all $w \in \mathcal{W} \subset \Re^1$, $\gamma(\lambda) = \partial\beta^{\mathrm{lr}}(\lambda)/\partial\lambda$ equals

$$\begin{aligned}\gamma(\lambda) = & \int_w \int_x \int_z \frac{\partial g}{\partial w}(F^{-1}_{W|Z}(F_{W_\lambda|Z}(\lambda x + (1-\lambda^2)^{1/2} w|z)|z), x, z) \\ & \times \frac{1}{\mathrm{d}F_{W|Z}(F^{-1}_{W|Z}(F_{W_\lambda|Z}(\lambda x + (1-\lambda^2)^{1/2} w|z)|z)|z)} \\ & \qquad \times \frac{\partial F_{W_\lambda|Z}(\lambda x + (1-\lambda^2)^{1/2} w|z)}{\partial\lambda} \cdot \mathrm{d}F_{W,X,Z}(w, x, z).\end{aligned}$$

In order to analyze $\gamma(\lambda)$ we begin by deriving an alternative expression for $\partial F_{W_\lambda|Z}(\lambda x + (1-\lambda^2)^{1/2} w|z)/\partial\lambda$. Observe that

$$F_{W_\lambda|Z}(\lambda x + (1-\lambda) w|z) = \Pr(W_\lambda \leq \lambda x + (1-\lambda^2)^{1/2} w|Z = z) = \Pr(W_\lambda \in A|Z = z),$$

where the set A is given by

$$\begin{aligned}A &= \left\{(W, X) : \lambda X + (1-\lambda^2)^{1/2} W \leq \lambda x + (1-\lambda^2)^{1/2} w,\ X \in \mathbb{R}^1,\ W \in \mathbb{R}^1\right\} \\ &= \left\{(W, X) : W \leq \frac{\lambda x + (1-\lambda^2)^{1/2} w - \lambda X}{(1-\lambda^2)^{1/2}},\ X \in \mathbb{R}^1\right\},\end{aligned}$$

which provides the limits of integration required to calculate $F_{W_\lambda|Z}(\lambda x + (1-\lambda^2)^{1/2} w|z)$:

$$F_{W_\lambda|Z}(\lambda x + (1-\lambda^2)^{1/2} w|z) = \int_{s=-\infty}^{s=\infty} \int_{v=-\infty}^{v=\frac{\lambda x + (1-\lambda^2)^{1/2} w - \lambda s}{(1-\lambda^2)^{1/2}}} \mathrm{d}F_{W,X|Z}(v, s|z). \tag{1.127}$$

By the Chain Rule we have, for $g(\lambda) = \frac{\lambda x + (1-\lambda^2)^{1/2} w - \lambda s}{(1-\lambda^2)^{1/2}}$,

$$\frac{\partial F_{W_\lambda|Z}(\lambda x + (1-\lambda^2)^{1/2} w|z)}{\partial\lambda} = \frac{\partial F_{W_\lambda|Z}(\lambda x + (1-\lambda^2)^{1/2} w|z)}{\partial g(\lambda)} g'(\lambda; s).$$

where

$$g'(\lambda; s) = \frac{1 + \lambda^2(1-\lambda^2)^{-1}}{(1-\lambda^2)^{1/2}}(x - s).$$

Applying the 'Second Fundamental Theorem of Calculus' to differentiate (1.127) with respect to $g(\lambda)$ gives

$$\frac{\partial F_{W_\lambda|Z}(\lambda x+(1-\lambda^2)^{1/2}w|z)}{\partial g(\lambda)} = \int_{s=-\infty}^{s=\infty} \mathrm{d}F_{W,X|Z}\left(\frac{\lambda x+(1-\lambda^2)^{1/2}w-\lambda s}{(1-\lambda^2)^{1/2}}, s|z\right),$$

and hence

$$\frac{\partial F_{W_\lambda|Z}(\lambda x+(1-\lambda^2)^{1/2}w|z)}{\partial \lambda} = \int_{s=-\infty}^{s=\infty} g'(\lambda;s)\,\mathrm{d}F_{W,X|Z}\left(\frac{\lambda x+(1-\lambda^2)^{1/2}w-\lambda s}{(1-\lambda^2)^{1/2}}, s|z\right). \tag{1.128}$$

Substituting (1.128) into the expression for $\gamma(\lambda)$ given above yields

$$\begin{aligned}\gamma(\lambda) &= \int_w\int_x\int_z \frac{\partial g}{\partial w}(F^{-1}_{W|Z}(F_{W_\lambda|Z}(\lambda x+(1-\lambda^2)^{1/2}w|z)|z),x,z)\\ &\times \frac{1}{\mathrm{d}F_{W|Z}(F^{-1}_{W|Z}(F_{W_\lambda|Z}(\lambda x+(1-\lambda^2)^{1/2}w|z)|z)|z)}\\ &\quad\times\left\{\int_{s=-\infty}^{s=\infty} g'(\lambda;s)\,dF_{W,X|Z}\left(\frac{\lambda x+(1-\lambda^2)^{1/2}w-\lambda s}{(1-\lambda^2)^{1/2}}, s|z\right)\right\}\cdot \mathrm{d}F_{W,X,Z}(w,x,z),\end{aligned}$$

which, after setting $\lambda=0$, gives γ_0 equal to

$$\begin{aligned}\gamma_0 &= \int_w\int_x\int_z \frac{\partial g}{\partial w}(w,x,z)\cdot\frac{1}{\mathrm{d}F_{W|Z}(w|z)}\times\left\{\int_{s=-\infty}^{s=\infty}(x-s)\,\mathrm{d}F_{W,X|Z}(w,s|z)\right\}\cdot\mathrm{d}F_{X,W,Z}(x,w,z)\\ &= \int_w\int_x\int_z \frac{\partial g}{\partial w}(w,x,z)\left\{\int_{s=-\infty}^{s=\infty}(x-s)\,\mathrm{d}F_{X|W,Z}(s|w,z)\right\}\cdot\mathrm{d}F_{X,W,Z}(x,w,z)\\ &= \mathbb{E}\left[\frac{\partial g}{\partial w}(W,X,Z)\cdot(X-\mathbb{E}[X|W,Z])\right],\end{aligned}$$

as claimed in the statement of the Theorem.

The weighted average derivative representation (5.23) follows by iterated expectations and Lemma A.17 above. By iterated expectations we have

$$\begin{aligned}\gamma_0 &= \mathbb{E}\left[\frac{\partial g}{\partial w}(W,X,Z)\cdot(X-\mathbb{E}[X|W,Z])\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{\partial g}{\partial w}(W,X,Z)\cdot(X-\mathbb{E}[X|W,Z])\,|W,Z\right]\right]\\ &= \mathbb{E}\left[Cov\left(\frac{\partial g}{\partial w}(W,X,Z),X|W,Z\right)\right].\end{aligned}$$

Applying Lemma A.17 with $Q=(W,Z')'$ then gives the desired result.

A.5 Proof to Theorem 5.1

First we establish some preliminary notation. As in the main text we have $K_\sigma(v - V_i) = \sigma^{-K+2} K\left(\frac{v-V_i}{b}\right)$. We let C be a generic constant which may vary from row-to-row. We define

$$\begin{aligned}
h_1(w,x,z) &= f_{W,X,Z}(w,x,z)\,\mathbb{E}[Y|W=w,X=x,Z=z] \\
h_2(w,x,z) &= f_{W,X,Z}(w,x,z) \\
h_3(w,z) &= f_{W,Z}(w,z)\,\mathbb{E}[X|W=w,Z=z] \\
h_4(w,z) &= f_{W,Z}(w,z)
\end{aligned}$$

and their corresponding estimates

$$\begin{aligned}
\widehat{h}_1(w,x,z) &= \frac{1}{N}\sum_{i=1}^{N} Y_i K_\sigma(w-W_i, x-X_i, z-Z_i) \\
\widehat{h}_2(w,x,z) &= \frac{1}{N}\sum_{i=1}^{N} K_\sigma(w-W_i, x-X_i, z-Z_i) \\
\widehat{h}_3(w,z) &= \frac{1}{N}\sum_{i=1}^{N} X_i K_\sigma(w-W_i, z-Z_i) \\
\widehat{h}_4(w,z) &= \frac{1}{N}\sum_{i=1}^{N} K_\sigma(w-W_i, z-Z_i).
\end{aligned}$$

The two regression functions entering the estimand are ratios of these objects, i.e.,

$$g(w,x,z) = \frac{h_1(w,x,z)}{h_2(w,x,z)}, \quad \widehat{g}(w,x,z) = \frac{\widehat{h}_1(w,x,z)}{\widehat{h}_2(w,x,z)}$$

and

$$m(w,z) = \frac{h_3(w,z)}{h_4(w,z)}, \quad \widehat{m}(w,z) = \frac{\widehat{h}_3(w,z)}{\widehat{h}_4(w,z)}.$$

We seek to characterize the large sample properties of the statistic

$$\widehat{\gamma} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \widehat{g}(W_i,X_i,Z_i)}{\partial w} \times (X_i - \widehat{m}(W_i,Z_i)).$$

In order to derive the large sample properties of $\widehat{\gamma}$ we begin by linearizing. Using the identity

$$\widehat{a}\widehat{b} = ab + a(\widehat{b}-b) + b(\widehat{a}-a) + (\widehat{a}-a)(\widehat{b}-b) \tag{1.129}$$

we have

$$\begin{aligned}
\widehat{\gamma} =\; & \frac{1}{N}\sum_{i=1}^{N} \frac{\partial g(W_i,X_i,Z_i)}{\partial w} \times (X_i - m(W_i,Z_i)) && (1.130) \\
& -\frac{1}{N}\sum_{i=1}^{N} \frac{\partial g(W_i,X_i,Z_i)}{\partial w} \times (m(W_i,Z_i) - \widehat{m}(W_i,Z_i)) && (1.131) \\
& +\frac{1}{N}\sum_{i=1}^{N} (X_i - m(W_i,Z_i))\left(\frac{\partial g(W_i,X_i,Z_i)}{\partial w} - \frac{\partial \widehat{g}(W_i,X_i,Z_i)}{\partial w}\right) && (1.132) \\
& -\frac{1}{N}\sum_{i=1}^{N} \left(\frac{\partial g(W_i,X_i,Z_i)}{\partial w} - \frac{\partial \widehat{g}(W_i,X_i,Z_i)}{\partial w}\right) \times (m(W_i,Z_i) - \widehat{m}(W_i,Z_i)) && (1.133)
\end{aligned}$$

Where (1.130) is the main term, (1.131) and (1.132) are correction terms which account for sampling error in $\widehat{m}(w,z)$ and $\widehat{g}(w,x,z)$ respectively and (1.133) is an $o_p(1)$ remainder term.

In order to obtain an asymptotically linear representation of $\widehat{\gamma}$ we express (1.131) and (1.132) as averages and verify that (1.133) is asymptotically negligible. We analyze each term in the above expansion separately

A.5.1 Analysis of (1.131) – correction term for nonparametric estimation of $m(w,z)$

We first expand (1.131) using the identity

$$\begin{aligned} \frac{\widehat{a}}{\widehat{b}} - \frac{a}{b} &= \frac{1}{b}(\widehat{a}-a) - \frac{a}{b^2}(\widehat{b}-b) + \frac{a}{b^2\widehat{b}}(\widehat{b}-b)^2 - \frac{a}{b\widehat{b}}(\widehat{a}-a)(\widehat{b}-b) \qquad (1.134) \\ &= \frac{1}{b}\left(\widehat{a} - \frac{a}{b}\widehat{b}\right) + \frac{a}{b^2\widehat{b}}(\widehat{b}-b)^2 - \frac{a}{b\widehat{b}}(\widehat{a}-a)(\widehat{b}-b) \end{aligned}$$

Using the second line of (1.134) we have

$$\begin{aligned} m(w,z) - \widehat{m}(w,z) &= \frac{1}{h_4(w,z)}(\widehat{h}_3(w,z) - m(w,z)\widehat{h}_4(w,z)) \\ &\quad + \frac{h_3(w,z)}{h_4(w,z)^2\,\widehat{h}_4(w,z)}(\widehat{h}_4(w,z) - h_4(w,z))^2 \\ &\quad - \frac{1}{h_4(w,z)\,\widehat{h}_4(w,z)}(\widehat{h}_3(w,z) - h_3(w,z))(\widehat{h}_4(w,z) - h_4(w,z)). \end{aligned}$$

Substituting this into (1.131) yields $-\frac{1}{N}\sum_{i=1}^{N}\frac{\partial g(W_i,X_i,Z_i)}{\partial w} \times (m(W_i,Z_i) - \widehat{m}(W_i,Z_i))$ equal to

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{\partial g(W_i,X_i,Z_i)}{\partial w} \times \frac{1}{h_4(W_i,Z_i)}(\widehat{h}_3(W_i,Z_i) - m(W_i,Z_i)\widehat{h}_4(W_i,Z_i)) \qquad (1.135)$$

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{\partial g(W_i,X_i,Z_i)}{\partial w} \times \frac{h_3(W_i,Z_i)}{h_4(W_i,Z_i)^2\,\widehat{h}_4(W_i,Z_i)}(\widehat{h}_4(W_i,Z_i) - h_4(W_i,Z_i))^2 \qquad (1.136)$$

$$\begin{aligned} &+\frac{1}{N}\sum_{i=1}^{N}\frac{\partial g(W_i,X_i,Z_i)}{\partial w} \qquad (1.137) \\ &\times\frac{1}{h_4(W_i,Z_i)\,\widehat{h}_4(W_i,Z_i)}(\widehat{h}_3(W_i,Z_i) - h_3(W_i,Z_i))(\widehat{h}_4(W_i,Z_i) - h_4(W_i,Z_i)). \end{aligned}$$

The first term (1.135) can be expressed as an average, while the second and third terms, (1.136) and (1.137), are $o_p(1)$ remainder terms.

We first analyze (1.135), substitution of $h_4(W_i,Z_i)$, $\widehat{h}_3(W_i,Z_i)$ and $\widehat{h}_4(W_i,Z_i)$ gives

$$-\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\partial g(W_i,X_i,Z_i)}{\partial w}\frac{1}{f_{W,Z}(W_i,Z_i)}(X_j - m(W_i,Z_i))K_\sigma(W_i-W_j, Z_i-Z_j),$$

which is a V-statistic of order two with kernel

$$a(V_i,V_j) = -\frac{\partial g(W_i,X_i,Z_i)}{\partial w}\frac{1}{f_{W,Z}(W_i,Z_i)}(X_j - m(W_i,Z_i))K_\sigma(W_i-W_j, Z_i-Z_j), \qquad (1.138)$$

where $v = (w,x,z')'$. For some sequence $\sigma_N \downarrow 0$ let $a_N(V_i,V_j)$ be equal to (1.138) with $\sigma = \sigma_N$. Define

$$a_{iN} = a_{iN}(V_i) = \mathbb{E}[a_N(V_i,\widetilde{V}_j)], \quad a_{jN} = a_{jN}(V_j) = \mathbb{E}[a_N(\widetilde{V}_i,V_j)],$$

where expectations are taken with respect to the marginal distributions of the variables with a '$\sim$' above them. Also let $\overline{a}_N = \mathbb{E}[a_{iN}(V_i)] = \mathbb{E}[a_{jN}(V_j)]$.

If $\mathbb{E}[a(V_i,V_j)^2] < \infty$ and the data are i.i.d. we can apply the V-statistic Projection Theorem (e.g., Newey and McFadden 1994, Lemma 8.4, p. 2201) to show that, under appropriate conditions on the sequence $\sigma_N \downarrow 0$,

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_N(V_i,V_j) - \left\{\frac{1}{N}\sum_{i=1}^{N} a_{iN}(V_i) - \frac{1}{N}\sum_{j=1}^{N} a_{jN}(V_j)\right\} + \overline{a}_N = o_p(1/\sqrt{N}).$$

We show below, again under appropriate conditions on the sequence $\sigma_N \downarrow 0$, that only the $\sum_{i=1}^{N} a_{jN}(V_j)/N$ term will contribute to the asymptotic distribution of $\widehat{\gamma}$. This means that the asymptotic sampling distribution of (1.135) depends only upon $\sum_{i=1}^{N} a_{jN}(V_j)/\sqrt{N}$ (the large sample behavior of which we can be characterized using a central limit theorem).

Verifying $\mathbb{E}[a(V_i,V_j)^2] < \infty$ (as required for V-statistic projection) In order to apply the V-statistic Projection Theorem we need to demonstrate that $\mathbb{E}[a(V_i,V_j)^2] < \infty$. Since $\frac{\partial g(W_i,X_i,Z_i)}{\partial w}\frac{1}{f_{W,Z}(W_i,Z_i)}$ is bounded by the requirement that $f_{W,Z}(W_i,Z_i)$ is bounded below by zero and the boundedness of $\frac{\partial g(W_i,X_i,Z_i)}{\partial w}$ (Assumptions **X** and **X**) we need only show that

$$\mathbb{E}[(X_j - m(W_i,Z_i))^2 K_\sigma(W_i - W_j, Z_i - Z_j)^2] < \infty.$$

This condition is follows from the fact that

$$\mathbb{E}[X_j^2 K_\sigma(W_i - W_j, Z_i - Z_j)^2] < \infty$$
$$|\mathbb{E}[X_j m(W_i,Z_i) K_\sigma(W_i - W_j, Z_i - Z_j)^2]| < \infty$$
$$\mathbb{E}[m(W_i,Z_i)^2 K_\sigma(W_i - W_j, Z_i - Z_j)^2] < \infty.$$

Each of these conditions follows from the assumptions of the theorem. First we have

$$\begin{aligned}
\mathbb{E}[X_j^2 K_\sigma(W_i - W_j, Z_i - Z_j)^2] &= \mathbb{E}_{W_i,Z_i}\left[\int \mathbb{E}\left[X_j^2 | W_j = w, Z_j = z\right] K_\sigma(W_i - w, Z_i - z)^2\right. \\
&\quad \left. \times f_{W,Z}(w,z)\, dwdz \right] \\
&= \mathbb{E}_{W_i,Z_i}\left[\int \mathbb{E}\left[X_j^2 | W_j = W_i - \sigma r, Z_j = Z_i - \sigma t\right] K(r,t)^2\right. \\
&\quad \left. \times f_{W,Z}(W_i - \sigma r, Z_i - \sigma t)\, drdt \right] \\
&\leq C\mathbb{E}[X_j^2] < \infty
\end{aligned}$$

by boundedness of $\mathbb{E}\left[X^2\right]$ and the kernel (Assumptions **X** and **X**). The second equality above follows from a change-of-variables with $r = (W_i - w)/\sigma$ and $t = (Z_i - z)/\sigma$ and a Jacobian of $\sigma^{-(K+1)}$; we make repeated use of this type of change in variables. We also have

$$\begin{aligned}
\left|\mathbb{E}[X_j m(W_i,Z_i) K_\sigma(W_i - W_j, Z_i - Z_j)^2]\right| &\leq \mathbb{E}_{W_i,Z_i}[|m(W_i,Z_i)| \\
&\quad \times \int |m(w,z)| K_\sigma(W_i - w, Z_i - z)^2 f_{W,Z}(w,z)\, dwdz] \\
&\leq \mathbb{E}_{W_i,Z_i}[|m(W_i,Z_i)|]\,\mathbb{E}_{W_j,Z_j}[|m(W_j,Z_j)|] < \infty
\end{aligned}$$

by Assumptions **X to X**. Finally we require that

$$\mathbb{E}[m(W_i,Z_i)^2 K_\sigma(W_i - W_j, Z_i - Z_j)^2] < C\mathbb{E}[m(W_i,Z_i)^2] < \infty,$$

which follows by, again, Assumptions **X to X**.

V-statistic projection for (1.135) Calculating the V-statistic projection associated with (1.135) we get:

$$\frac{1}{N^2}\sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{N} a_N\left(V_i,V_j\right)=\frac{1}{N}\sum\nolimits_{i=1}^{N} a_{iN}\left(V_i\right)+\frac{1}{N}\sum\nolimits_{j=1}^{N} a_{jN}\left(V_j\right)+o_p(1/\sqrt{N})$$

where

$$a_{iN}\left(V_i\right)=-\int\frac{\partial g\left(W_i,X_i,Z_i\right)}{\partial w}\frac{1}{f_{W,Z}\left(W_i,Z_i\right)}(x-m\left(W_i,Z_i\right))K_{\sigma_N}\left(W_i-w,Z_i-z\right)f_{W,X,Z}\left(w,x,z\right)dwdxdz,$$

and

$$\begin{aligned}
a_{jN}\left(V_j\right) &= -\int\frac{\partial g\left(w,x,z\right)}{\partial w}\frac{1}{f_{W,Z}\left(w,z\right)}(X_j-m\left(w,z\right))K_{\sigma_N}\left(w-W_j,z-Z_j\right)f_{W,X,Z}\left(w,x,z\right)dwdxdz\\
&= -\int\frac{\partial g\left(w,x,z\right)}{\partial w}(X_j-m\left(w,z\right))K_{\sigma_N}\left(w-W_j,z-Z_j\right)f_{X|W,Z}\left(x|w,z\right)dwdxdz\\
&= -\int\mathbb{E}\left[\frac{\partial g\left(W,X,Z\right)}{\partial w}|W=w,Z=z\right](X_j-m\left(w,z\right))K_{\sigma_N}\left(w-W_j,z-Z_j\right)dwdz.
\end{aligned}$$

Observe that a_{iN} and a_{jN} are triangular arrays. We now demonstrate that only a_{jN} contributes to the asymptotic sampling distribution of $\widehat{\gamma}$.

First consider a_{jN}, we have $\left|\sqrt{N}\mathbb{E}[a_{jN}]\right|$ equal to

$$\begin{aligned}
\left|\sqrt{N}\mathbb{E}\left[a_{jN}\left(V_j\right)\right]\right| &= |\sqrt{N}\mathbb{E}_{W_j,X_j,Z_j}[\int\mathbb{E}\left[\frac{\partial g\left(W,X,Z\right)}{\partial w}|W=w,Z=z\right]\\
&\quad\times(X_j-m\left(w,z\right))K_{\sigma_N}\left(w-W_j,z-Z_j\right)dwdz]|\\
&= |\sqrt{N}\mathbb{E}_{W_j,X_j,Z_j}[\int\mathbb{E}\left[\frac{\partial g\left(W,X,Z\right)}{\partial w}|W=W_j+r\sigma_N,Z=Z_j+t\sigma_N\right]\\
&\quad\times(X_j-m\left(W_j+r\sigma_N,Z_j+t\sigma_N\right))K\left(r,t\right)drdt]|\\
&= |\sqrt{N}\int\mathbb{E}_{W_j,X_j,Z_j}[\mathbb{E}\left[\frac{\partial g\left(W,X,Z\right)}{\partial w}|W=W_j+r\sigma_N,Z=Z_j+t\sigma_N\right]\\
&\quad\times(X_j-m\left(W_j+r\sigma_N,Z_j+t\sigma_N\right))]K\left(r,t\right)drdt|\\
&= |\sqrt{N}\int\mathbb{E}_{W_j,Z_j}[\mathbb{E}\left[\frac{\partial g\left(W,X,Z\right)}{\partial w}|W=W_j+r\sigma_N,Z=Z_j+t\sigma_N\right]\\
&\quad\times(m\left(W_j,Z_j\right)-m\left(W_j+r\sigma_N,Z_j+t\sigma_N\right))]K\left(r,t\right)drdt|,
\end{aligned}$$

where the last line follows from an application of the law of iterated expectations. In order to calculate the rate of convergence of $\left|\sqrt{N}\mathbb{E}\left[a_{jN}\left(V_j\right)\right]\right|$ toward zero we replace $m\left(W_j,Z_j\right)-m\left(W_j+r\sigma_N,Z_j+t\sigma_N\right)$ with its $(S-1)^{th}$ order multivariate Taylor expansion in r and t about zero. This expansion is

$$\begin{aligned}
m\left(W_i+r\sigma_N,Z_i+t\sigma_N\right)-m\left(W_i,Z_i\right) &= \sum_{s=1}^{S-1}\frac{\sigma_N^{s(K+1)}}{s!}\left(r\frac{\partial}{\partial w}+t'\frac{\partial}{\partial z}\right)^s m\left(W_i,Z_i\right) \qquad (1.139)\\
&\quad+\sigma_N^{S(K+1)}\frac{1}{S!}\left(r\frac{\partial}{\partial w}+t'\frac{\partial}{\partial z}\right)^S m\left(W_i-\overline{r}\sigma_N,Z_i-\overline{t}\sigma_N\right),
\end{aligned}$$

where $\bar{r}$ and $\bar{t}$ are mean values between zero and r and t (c.f., Serfling 1980, p. 44). Since by Assumptions **X, X and X** we have $m(\cdot)$ differentiable with an S^{th} derivative bounded on $\mathbb{R}^{K+1}$, $g(\cdot)$ differentiable with an S^{th} derivative bounded on $\mathbb{R}^{K+2}$ (and hence the S^{th} derivative of $\mathbb{E}\left[\frac{\partial g(W,X,Z)}{\partial w}|W=w,Z=z\right]$ bounded on $\mathbb{R}^{K+1}$), and a kernel of order S, we conclude that

$$\left|\sqrt{N}\mathbb{E}\left[a_{jN}(V_j)\right]\right| \leq C\sqrt{N}\sigma_N^{S(K+1)},$$

and hence we conclude that $\sqrt{N}\mathbb{E}\left[a_{jN}(V_j)\right] = o_p(1)$ for $\sigma_N \downarrow 0$ such that $\sqrt{N}\sigma_N^{S(K+1)} \to 0$.

In order to apply the Lindeberg-Feller central limit theorem to $\sum_{i=1}^{N} a_{jN}(V_j)/\sqrt{N}$ we need to verify that the upper bound of $a_{jN}(V_j)^2$ has a finite expected value. By repeated application of the triangle inequality (TI) and a change of variables we have

$$\begin{aligned}
|a_{jN}(V_j)| &\leq C\int\left|\mathbb{E}\left[\frac{\partial g(W,X,Z)}{\partial w}|W=w,Z=z\right](X_j - m(w,z))K_{\sigma_N}(w-W_j,z-Z_j)\right|dwdz \\
&\leq C\int\left|\mathbb{E}\left[\frac{\partial g(W,X,Z)}{\partial w}|W=w,Z=z\right]X_jK_{\sigma_N}(w-W_j,z-Z_j)\right|dwdz \\
&\quad +C\int\left|\mathbb{E}\left[\frac{\partial g(W,X,Z)}{\partial w}|W=w,Z=z\right]m(w,z)K_{\sigma_N}(w-W_j,z-Z_j)\right|dwdz \\
&= C\int\left|\mathbb{E}\left[\frac{\partial g(W,X,Z)}{\partial w}|W=W_j+r\sigma_N,Z=Z_j+t\sigma_N\right]X_jK(r,t)\right|drdt \\
&\quad +C\int\left|\mathbb{E}\left[\frac{\partial g(W,X,Z)}{\partial w}|W=W_j+r\sigma_N,Z=Z_j+t\sigma_N\right]m(W_j+r\sigma_N,Z_j+t\sigma_N)K(r,t)\right|drdt \\
&\leq C_1\left|\mathbb{E}\left[\frac{\partial g(W,X,Z)}{\partial w}|W=W_j,Z=Z_j\right]X_j\right| \\
&\quad +C_2\left|\mathbb{E}\left[\frac{\partial g(W,X,Z)}{\partial w}|W=W_j,Z=Z_j\right]m(W_j,Z_j)\right|,
\end{aligned}$$

where the last line follows by the fact that $K(r,t)$ is maximal at $r=s=t=0$ and the integral can be interpreted as an 'expectation' where $K(r,t)$ is a 'density' Since both of the terms on the right hand side have finite expected values we have $\mathbb{E}[a_{jN}(V_j)^2] < \infty$.

Calculating $\mathbb{E}[a_{jN}^2]$ we get

$$\begin{aligned}
\mathbb{E}[a_{jN}^2] &= \mathbb{E}_{W_j,X_j,Z_j}[\int\int\frac{\partial g(w,x,z)}{\partial w}\frac{\partial g(w^*,x^*,z^*)}{\partial w}(X_j - m(w,z))(X_j - m(w^*,z^*)) \\
&\quad \times K_{\sigma_N}(w-W_j,z-Z_j)K_{\sigma_N}(w^*-W_j,z^*-Z_j) \\
&\quad \times f_{X|W,Z}(x|w,z)f_{X|W,Z}(x^*|w^*,z^*)dwdxdzdw^*dx^*dz^*] \\
&= \mathbb{E}_{W_j,X_j,Z_j}[\int\int\mathbb{E}\left[\frac{\partial g(W,X,Z)}{\partial w}|W=w,Z=z\right]\mathbb{E}\left[\frac{\partial g(W,X,Z)}{\partial w}|W=w^*,Z=z^*\right]\times \\
&\quad (X_j - m(w,z))(X_j - m(w^*,z^*)K_{\sigma_N}(w-W_j,z-Z_j)K_{\sigma_N}(w^*-W_j,z^*-Z_j)dwdzdw^*dz^*] \\
&= \mathbb{E}_{W_j,X_j,Z_j}[\int\int\mathbb{E}\left[\frac{\partial g(W,X,Z)}{\partial w}|W=W_j+r\sigma_N,Z=Z_j+t\sigma_N\right]\times \\
&\quad \mathbb{E}\left[\frac{\partial g(W,X,Z)}{\partial w}|W=W_j+r^*\sigma_N,Z=Z_j+t^*\sigma_N\right]\times \\
&\quad (X_j - m(W_j+r\sigma_N,Z_j+t\sigma_N))(X_j - m(W_j+r^*\sigma_N,Z_j+t^*\sigma_N)K(r,t)K(r^*,t^*)drdtdr^*dt^*],
\end{aligned}$$

where the double integral inside the expectation converges to

$$\mathbb{E}\left[\frac{\partial g\left(W,X,Z\right)}{\partial w}|W=W_j,Z=Z_j\right]^2\left(X_j-m\left(W_j,Z_j\right)\right)^2$$

for $\sigma_N\downarrow 0$ (since $g_N\left(x_0\right)=\int K\left(r\right)g\left(w-r\sigma_N\right)dr\to g\left(w\right)\int K\left(r\right)dr$ and the kernel integrates to one. Therefore by Lebesgue's Dominated Convergence Theorem we have that

$$\lim_{N\to\infty}\mathbb{E}[a_{jN}^2]=\mathbb{E}_{W_j,X_j,Z_j}\left[\mathbb{E}\left[\frac{\partial g\left(W,X,Z\right)}{\partial w}|W=W_j,Z=Z_j\right]^2\left(X_j-m\left(W_j,Z_j\right)\right)^2\right]=AVar(a_{jN}^2).$$

Combining our results we have, for any sequence σ_N such that $\sqrt{N}\sigma_N^{S(K+1)}\downarrow 0$,

$$\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\left[a_{jN}\left(V_j\right)-\mathbb{E}\left[a_{jN}\left(V_j\right)\right]\right]=\frac{1}{\sqrt{N}}\sum_{j=1}^{N}a_{jN}\left(V_j\right)+o_p\left(1\right),$$

and hence by the Lindeberg-Feller central limit theorem (c.f., van der Vaart 1998, p. 20) that

$$\frac{1}{\sqrt{N}}\sum_{j=1}^{N}a_{jN}\left(V_j\right)\to\mathcal{N}(0,AVar(a_{jN}^2)). \tag{1.140}$$

We now demonstrate that $\sum_{i=1}^{N}a_{iN}\left(V_i\right)/\sqrt{N}$ is $o_p\left(1\right)$ for $\sqrt{N}\sigma_N^{S(K+1)}\downarrow 0$ with a negligible variance and hence does not contribute to the asymptotic sampling distribution of (1.135). We have, using iterated expectations, change of variables and the Taylor expansion given in (1.139) above,

$$\begin{aligned}
a_{iN}\left(V_i\right) &= -\int\frac{\partial g\left(W_i,X_i,Z_i\right)}{\partial w}\frac{1}{f_{W,Z}\left(W_i,Z_i\right)}(x-m\left(W_i,Z_i\right))K_{\sigma_N}\left(W_i-w,Z_i-z\right)\\
&\quad\times f_{W,X,Z}\left(w,x,z\right)dwdxdz\\
&= -\frac{\partial g\left(W_i,X_i,Z_i\right)}{\partial w}\frac{1}{f_{W,Z}\left(W_i,Z_i\right)}\mathbb{E}_{\widetilde{W},\widetilde{X},\widetilde{Z}}[(\widetilde{X}-m\left(W_i,Z_i\right))K_{\sigma_N}(W_i-\widetilde{W},Z_i-\widetilde{Z})]\\
&= -\frac{\partial g\left(W_i,X_i,Z_i\right)}{\partial w}\frac{1}{f_{W,Z}\left(W_i,Z_i\right)}\mathbb{E}_{\widetilde{W},\widetilde{Z}}[(m(\widetilde{W},\widetilde{Z})-m\left(W_i,Z_i\right))\times K_{\sigma_N}(W_i-\widetilde{W},Z_i-\widetilde{Z})]\\
&= -\int\frac{\partial g\left(W_i,X_i,Z_i\right)}{\partial w}\frac{1}{f_{W,Z}\left(W_i,Z_i\right)}(m(W_i-r\sigma_N,Z_i-t\sigma_N)-m\left(W_i,Z_i\right))\times\\
&\quad K(r,t)f_{W,Z}\left(W_i-r\sigma_N,Z_i-t\sigma_N\right)drdt\\
&= -[\int\frac{\partial g\left(W_i,X_i,Z_i\right)}{\partial w}\frac{1}{f_{W,Z}\left(W_i,Z_i\right)}(\sum_{s=1}^{S-1}\frac{\sigma_N^{s(K+1)}}{s!}\left(r\frac{\partial}{\partial w}+t'\frac{\partial}{\partial z}\right)^s m\left(W_i,Z_i\right)\\
&\quad+\sigma_N^{S(K+1)}\frac{1}{S!}\left(r\frac{\partial}{\partial w}+t'\frac{\partial}{\partial z}\right)^S m\left(W_i-\overline{r}\sigma_N,Z_i-\overline{t}\sigma_N\right))\times K(r,t)\\
&\quad\times f_{W,Z}\left(W_i-r\sigma_N,Z_i-t\sigma_N\right)drdt].
\end{aligned}$$

By the repeated application of triangle inequality (TI) we get

$$\begin{aligned}
|a_{iN}(V_i)| &\leq C\int|\frac{\partial g(W_i,X_i,Z_i)}{\partial w}\frac{1}{f_{W,Z}(W_i,Z_i)}(\sum_{s=1}^{S-1}\frac{\sigma_N^{s(K+1)}}{s!}\left(r\frac{\partial}{\partial w}+t'\frac{\partial}{\partial z}\right)^s m(W_i,Z_i) \\
&\quad +\sigma_N^{S(K+1)}\frac{1}{S!}\left(r\frac{\partial}{\partial w}+t'\frac{\partial}{\partial z}\right)^S m\left(W_i-\overline{r}\sigma_N,Z_i-\overline{t}\sigma_N\right)) \\
&\quad \times K(r,t)f_{W,Z}(W_i-r\sigma_N,Z_i-t\sigma_N)|drdt \\
&\leq C_1\int|\frac{\partial g(W_i,X_i,Z_i)}{\partial w}\frac{1}{f_{W,Z}(W_i,Z_i)}(\sum_{s=1}^{S-1}\frac{\sigma_N^{s(K+1)}}{s!}\left(r\frac{\partial}{\partial w}+t'\frac{\partial}{\partial z}\right)^s m(W_i,Z_i)\times \\
&\quad K(r,t)f_{W,Z}(W_i-r\sigma_N,Z_i-t\sigma_N)|drdt \\
&\quad +C_2\sigma_N^{S(K+1)}\int|\frac{\partial g(W_i,X_i,Z_i)}{\partial w}\frac{1}{f_{W,Z}(W_i,Z_i)}\frac{1}{S!}\left(r\frac{\partial}{\partial w}+t'\frac{\partial}{\partial z}\right)^S m\left(W_i-\overline{r}\sigma_N,Z_i-\overline{t}\sigma_N\right)\times \\
&\quad K(r,t)f_{W,Z}(W_i-r\sigma_N,Z_i-t\sigma_N)|drdt.
\end{aligned}$$

By virtue of the kernel being of order S and boundedness of $f_{W,X,Z}(W_i,X_i,Z_i)$, $g(W_i,X_i,Z_i)$, $m(W_i,Z_i)$ and their respective derivatives to order S (Assumptions **X, X and X**) we therefore have

$$|a_{iN}(V_i)|\leq C\sigma_N^{S(K+1)}.$$

Therefore the contribution of $a_{iN}(V_i)$ to the asymptotic variance of $\widehat{\gamma}$ will be negligible for bandwidth sequences such that $\sqrt{N}\sigma_N^{S(K+1)}\downarrow 0$ as claimed.

[REMAINDER TO BE COMPLETED]

References

[1] Newey, Whitney K. and Daniel McFadden. (1994). "Large sample estimation and hypothesis testing," *Handbook of Econometrics 4*: 2111 - 2245 (R.F. Engle & D.L. McFadden). Amsterdam: North Holland.

[2] Serfling, Robert J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: John Wiley & Sons, Inc.

[3] van der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge: Cambridge University Press.

A.6 Large sample distribution for AREs with discretely-valued inputs

[TO BE COMPLETED]